CS 224S/LING 281 Speech Recognition, Synthesis, and Dialogue

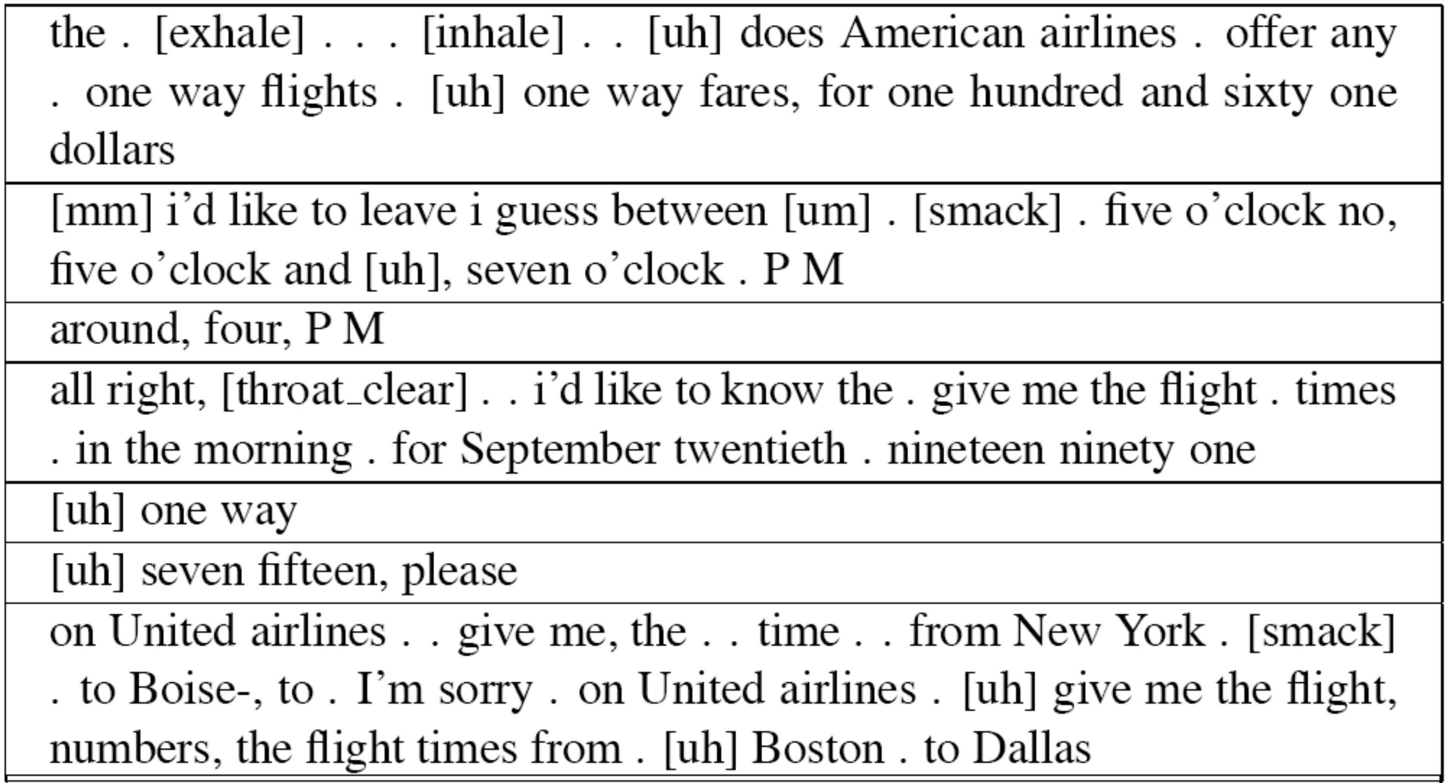
Dan Jurafsky

Lecture 17: Disfluencies

Outline

- Disfluencies
- Characteristics of disfluences
- Detecting disfluencies
- MDE bakeoff
- Fragments

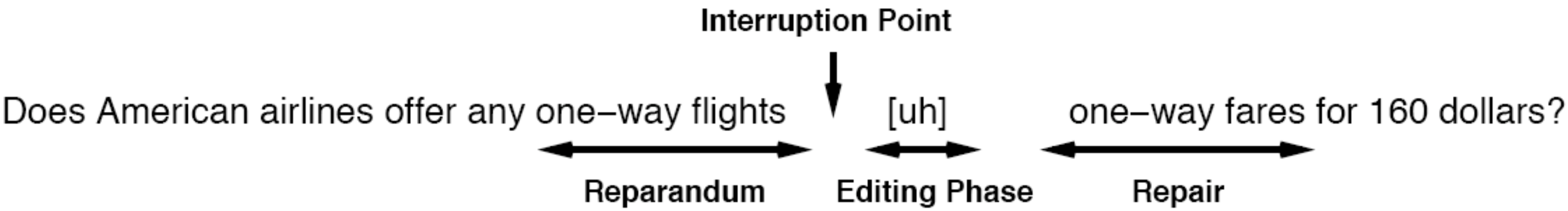
Disfluencies



the . [exhale] . . . [inhale] . . [uh] does American airlines . offer any . one way flights . [uh] one way fares, for one hundred and sixty one dollars
[mm] i'd like to leave i guess between [um] . [smack] . five o'clock no, five o'clock and [uh], seven o'clock . P M
around, four, P M
all right, [throat_clear] . . i'd like to know the . give me the flight . times . in the morning . for September twentieth . nineteen ninety one
[uh] one way
[uh] seven fifteen, please
on United airlines . . give me, the . . time . . from New York . [smack] . to Boise-, to . I'm sorry . on United airlines . [uh] give me the flight, numbers, the flight times from . [uh] Boston . to Dallas

Figure 9.5 Some sample spoken utterances from users interacting with the ATIS system.

Disfluencies:
standard terminology (Levelt)


Interruption Point

Does American airlines offer any one–way flights [uh] one–way fares for 160 dollars?

Reparandum

Editing Phase

Repair

- Reparandum: thing repaired
- Interruption point (IP): where speaker breaks off
- Editing phase (edit terms): uh, I mean, you know
- Repair: fluent continuation

Why disfluencies?

- Need to clean them up to get understanding
  - Does American airlines offer any ~~one-way flights [uh]~~ one-way fares for 160 dollars?
  - Delta leaving Boston seventeen twenty one arriving Fort Worth ~~twenty two~~ twenty one forty
- Might help in language modeling
  - Disfluencies might occur at particular positions (boundaries of clauses, phrases)
- Annotating them helps readability
- Disfluencies cause errors in nearby words

Counts (from Shriberg, Heeman)

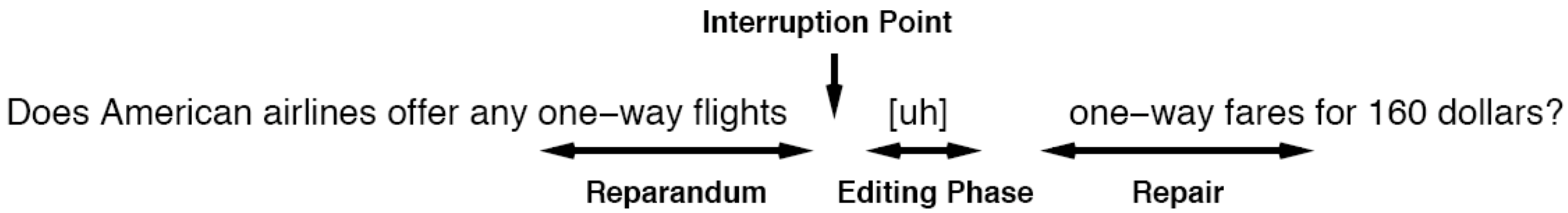
- Sentence disfluency rate
  - ATIS: 6% of sentences disfluent (10% long sentences)
  - Levelt human dialogs: 34% of sentences disfluent
  - Swbd: ~50% of multiword sentences disfluent
  - TRAINS: 10% of words are in reparandum or editing phrase
- Word disfluency rate
  - SWBD: 6%
  - ATIS: 0.4%
  - AMEX 13%
    - (human-human air travel)

Prosodic characteristics of disfluencies

- Nakatani and Hirschberg 1994
- Fragments are good cues to disfluencies
- Prosody:
  - Pause duration is shorter in disfluent silence than fluent silence
  - F0 increases from end of reparandum to beginning of repair, but only minor change
  - Repair interval offsets have minor prosodic phrase boundary, even in middle of NP:
    - Show me all n- | round-trip flights | from Pittsburgh | to Atlanta

Syntactic Characteristics of Disfluencies

- Hindle (1983)
- The repair often has same structure as reparandum
- Both are Noun Phrases (NPs) in this example:



Does American airlines offer any one–way flights [uh] one–way fares for 160 dollars?

Disfluencies and LM

- Clark and Fox Tree
- Looked at “um” and “uh”
  - “uh” includes “er” (“er” is just British non-rhotic dialect spelling for “uh”)
- Different meanings
  - Uh: used to announce minor delays
    - Preceded and followed by shorter pauses
  - Um: used to announce major delays
    - Preceded and followed by longer pauses

Um versus uh: delays
(Clark and Fox Tree)

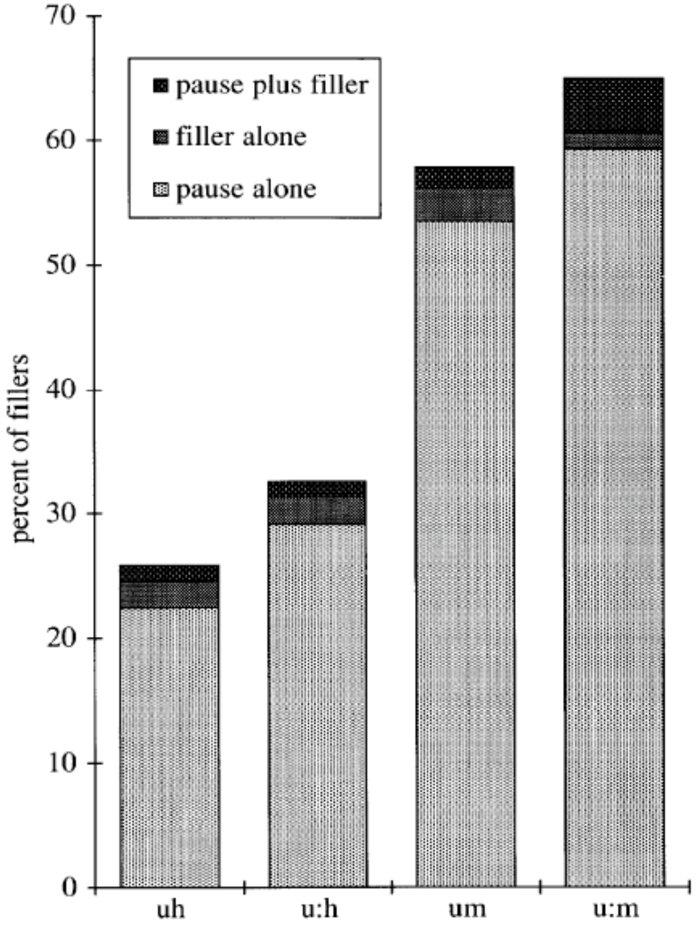


Fig. 1. Percent of fillers followed by delays (LL corpus).

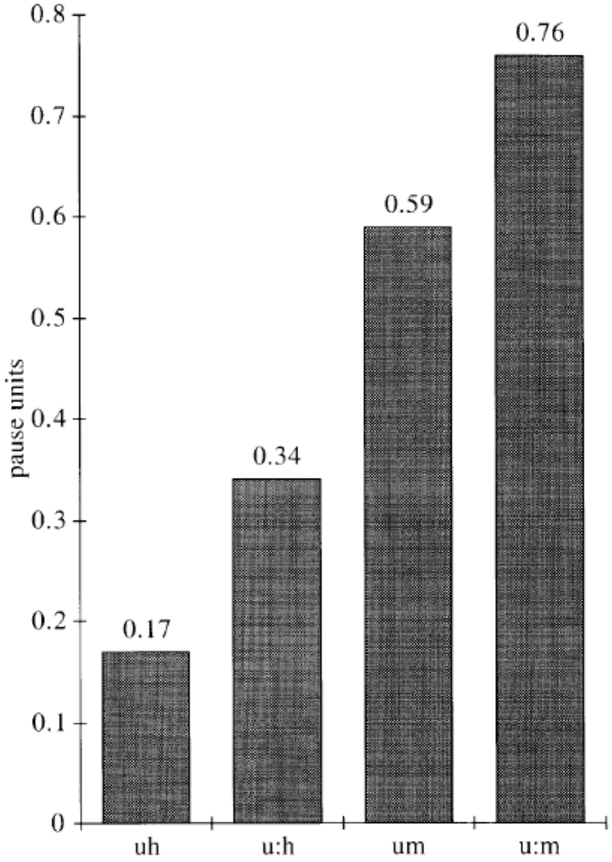


Fig. 2. Mean length of pauses after fillers (LL corpus).

Utterance Planning

- The more difficulty speakers have in planning, the more delays
- Consider 3 locations:
  - I: before intonation phrase: hardest
  - II: after first word of intonation phrase: easier
  - III: later: easiest
- And then uh somebody said, . [I] but um -- [II] don’t you think there’s evidence of this, in the twelfth - [III] and thirteenth centuries?

Delays at different points in phrase

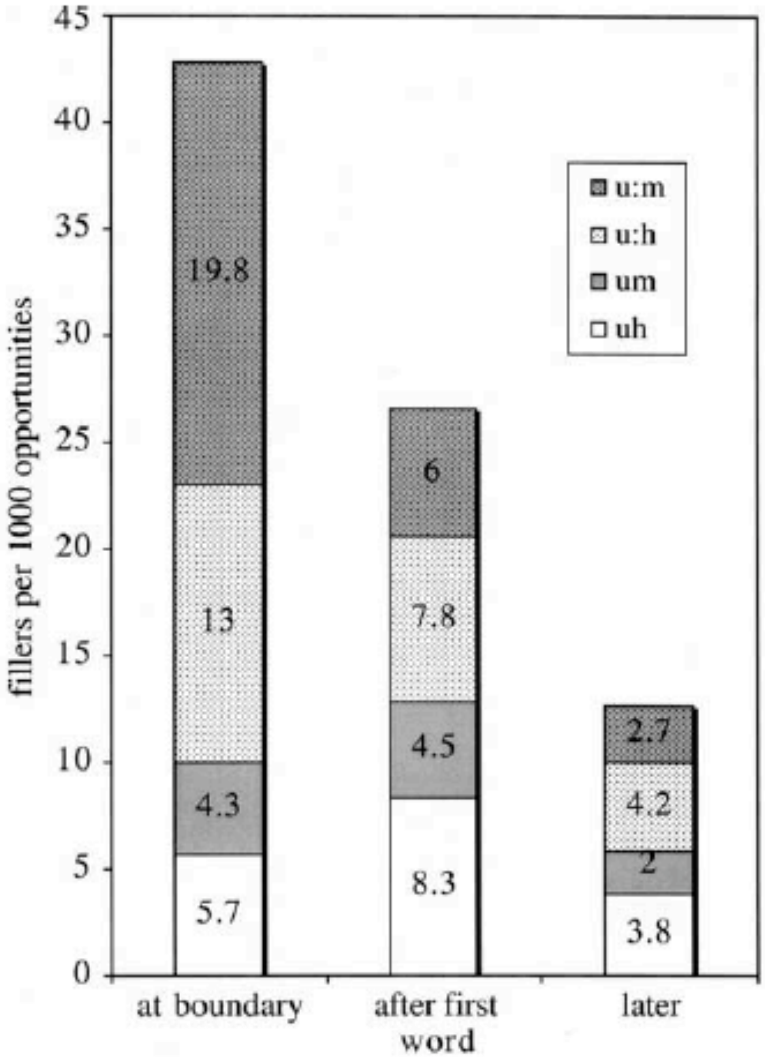


Fig. 5. Rates of *uh* and *um* at three position in tone units (LL corpus).

Disfluencies in language modeling

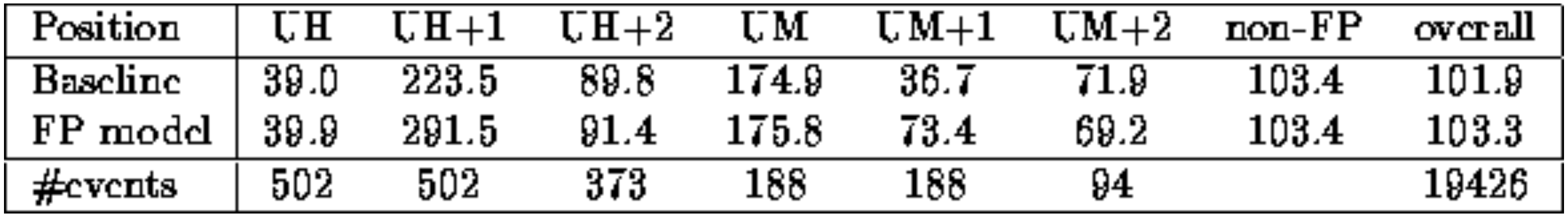
- Should we "clean up" disfluencies before training LM (i.e. skip over disfluencies?)
  - Filled pauses
    - Does United offer any [uh] one-way fares?
  - Repetitions
    - What what are the fares?
  - Deletions
    - Fly to Boston from Boston
  - Fragments (we'll come back to these)
    - I want fl- flights to Boston.

Does deleting disfluencies improve LM perplexity?

- Stolcke and Shriberg (1996)
- Build two LMs
  - Raw
  - Removing disfluencies
- See which has a higher perplexity on the data.
  - Filled Pause
  - Repetition
  - Deletion

Change in Perplexity when Filled Pauses (FP) are removed

- LM Perplexity goes up at following word:



Position	UH	UH+1	UH+2	UM	UM+1	UM+2	non-FP	overall
Baseline	39.0	223.5	89.8	174.9	36.7	71.9	103.4	101.9
FP model	39.9	291.5	91.4	175.8	73.4	69.2	103.4	103.3
#events	502	502	373	188	188	94		19426

- Removing filled pauses makes LM worse!!
- I.e., filled pauses seem to help to predict next word.
- Why would that be?

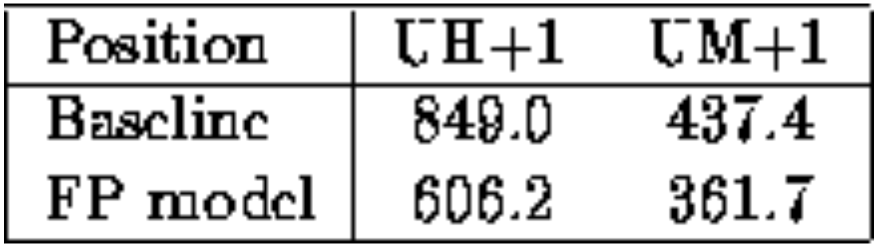
Stolcke and Shriberg 1996

Filled pauses tend to occur at clause boundaries

- Word before FP is end of previous clause; word after is start of new clause;
  - Best not to delete FP
- Some of the different things we’re doing [uh] there’s not time to do it all
- “there’s” is very likely to start a sentence
- So P(there’s|uh) is better estimate than P(there’s|doing)

Suppose we just delete medial FPs

- Experiment 2:
  - Parts of SWBD hand-annotated for clauses
  - Build FP-model by deleting only medial FPs
  - Now prediction of post-FP word (perplexity) improves greatly!
  - Siu and Ostendorf found same with “you know”



Position	UH+1	UM+1
Baseline	849.0	437.4
FP model	606.2	361.7

What about REP and DEL

- S+S built a model with “cleaned-up” REP and DEL
- Slightly lower perplexity
- But exact same word error rate (49.5%)
- Why?
  - Rare: only 2 words per 100
  - Doesn’t help because adjacent words are misrecognized anyhow!

Stolcke and Shriberg conclusions wrt LM and disfluencies

- Disfluency modeling purely in the LM probably won't vastly improve WER
- But
  - Disfluencies should be considered jointly with sentence segmentation task
  - Need to do better at recognizing disfluent words themselves
  - Need acoustic and prosodic features

WER reductions from modeling disfluencies+background events

- Rose and Riccardi (1999)

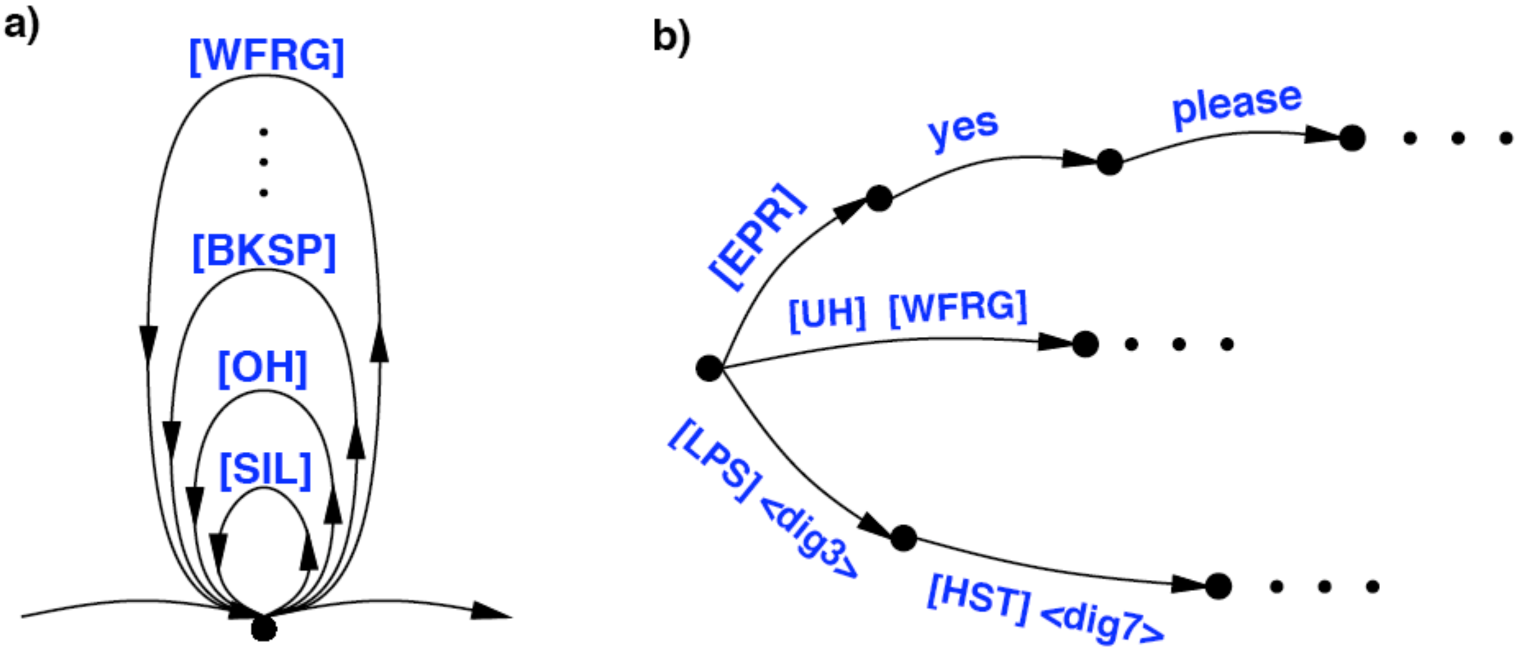


Figure 1: *a)* Inclusion of all labeled background events (LBEs) in a single "between–word" loop. *b)* Portion of phrase–based LM trained from LBE annotated text.

HMIHY Background events

- Out of 16,159 utterances:

Filled Pauses	7189
Word Fragments	1265
Hesitations	792
Laughter	163
Lipsmack	2171
Breath	8048
Non-Speech Noise	8834
Background Speech	3585
Operater Utt.	5112
Echoed Prompt	5353

Phrase-based LM

- “I would like to make a collect call”
- “a [wfrag]”
- <dig3> [brth] <dig3>
- “[brth] and I”

Rose and Riccardi (1999)
Modeling “LBEs” does help in WER

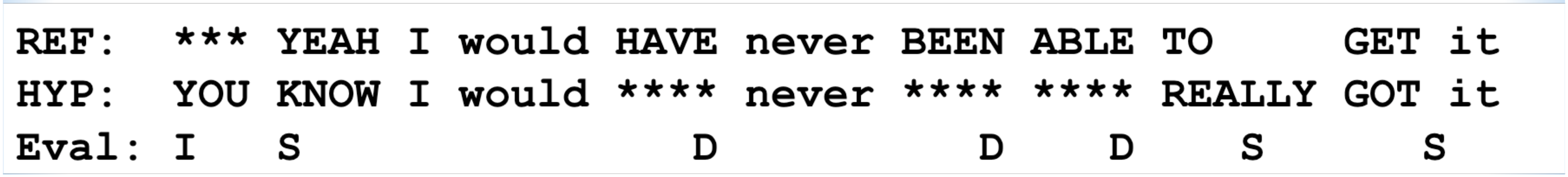
ASR Word Accuracy			
System Configuration		Test Corpora	
HMM	LM	Greeting (HMIHY)	Card Number
Baseline	Baseline	58.7	87.7
LBE	Baseline	60.8	88.1
LBE	LBE	60.8	89.8

More on location of FPs

- Peters: Medical dictation task
  - Monologue rather than dialogue
  - In this data, FPs occurred INSIDE clauses
  - Trigram PP after FP: 367
  - Trigram PP after word: 51
- Stolcke and Shriberg (1996b)
  - w_k FP w_{k+1}: looked at $P(w_{k+1}|w_k)$
  - Transition probabilities lower for these transitions than normal ones
- Conclusion:
  - People use FPs when they are planning difficult things, so following words likely to be unexpected/ rare/difficult

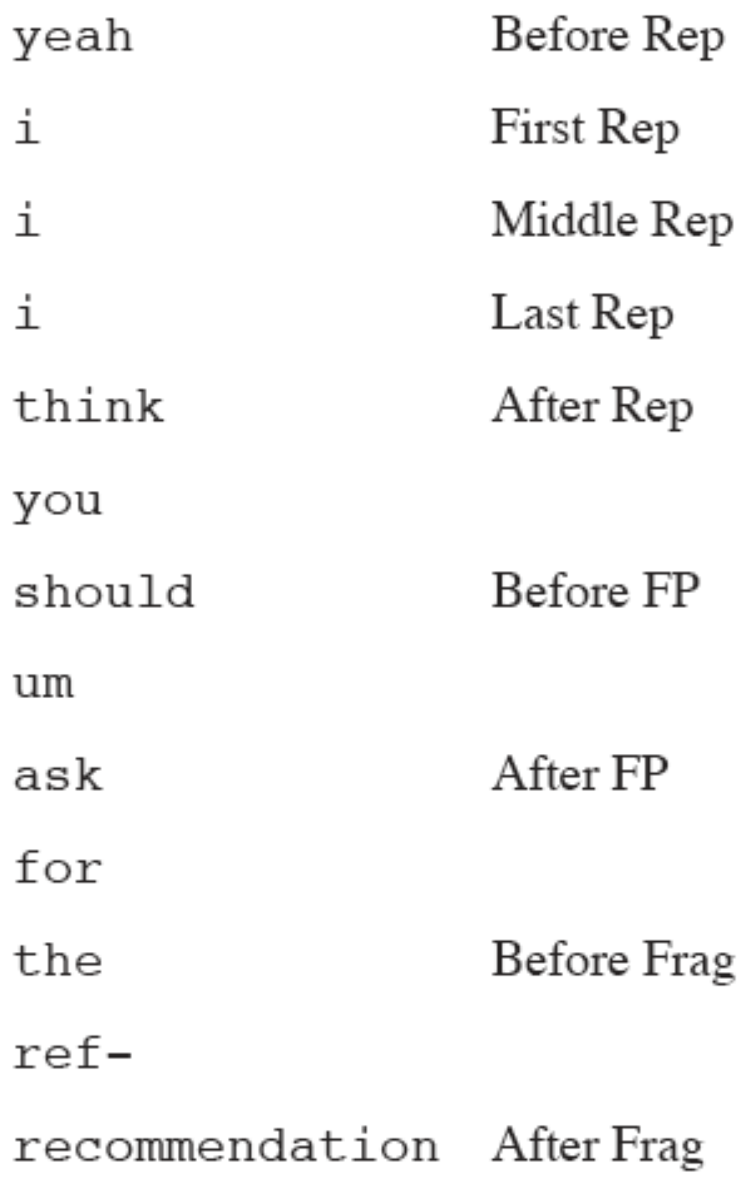
Goldwater study

- Study error rate in two recognizers
  - SRI/ICSI/UW RT-04 CTS system (Stolcke et al., 2006)
  - CU-HTK RT-04 CTS system (Evermann 2004a, 2005)
- On NIST RT-03 development set.
  - 36 telephone conversations, 72 speakers, 38477 reference words.
  - Metric: Individual Word Error Rate:



```
REF:  *** YEAH I would HAVE never BEEN ABLE TO     GET it
HYP:  YOU KNOW I would **** never **** **** REALLY GOT it
Eval: I   S               D          D    D    S      S
```

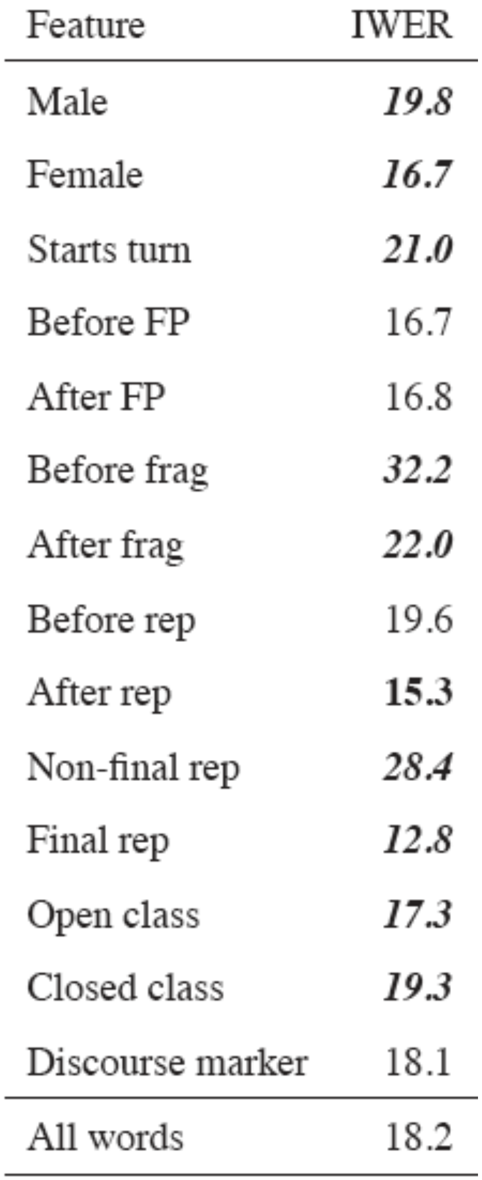
Coding disfluencies



yeah	Before Rep
i	First Rep
i	Middle Rep
i	Last Rep
think	After Rep
you	
should	Before FP
um	
ask	After FP
for	
the	Before Frag
ref-	
recommendation	After Frag

Other factors and (independent) results

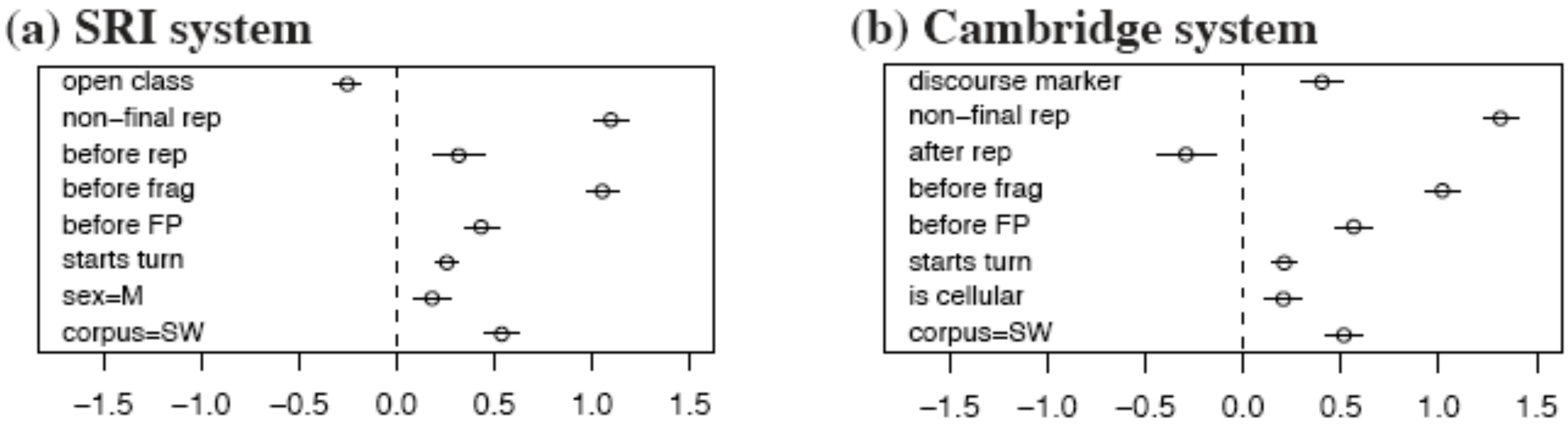
- Word class
  - Open
  - Function
  - Discourse marker
- Turn-initial
- Male
- Female
- Word length in phones
- LM unigram, trigram
- Prosodic features



Feature	IWER
Male	***19.8***
Female	***16.7***
Starts turn	***21.0***
Before FP	16.7
After FP	16.8
Before frag	***32.2***
After frag	***22.0***
Before rep	19.6
After rep	**15.3**
Non-final rep	***28.4***
Final rep	***12.8***
Open class	***17.3***
Closed class	***19.3***
Discourse marker	18.1
All words	18.2

Coefficient values in joint regression

- Left of dotted line -> reduces error
- Right of dotted line -> increases error



Detection of disfluencies

- Nakatani and Hirschberg
- Decision tree at w_i-w_j boundary
  - pause duration
  - Word fragments
  - Filled pause
  - Energy peak within w_i
  - Amplitude difference between w_i and w_j
  - F0 of w_i
  - F0 differences
  - Whether wi accented
- Results:
  - 78% recall/89.2% precision

Detection/Correction

- Bear, Dowding, Shriberg (1992)
- System 1:
- Hand-written pattern matching rules to find repairs
    - Look for identical sequences of words
    - Look for syntactic anomalies (“a the”, “to from”)
    - 62% precision, 76% recall
    - Rate of accurately correcting: 57%

Using Natural Language Constraints

- Gemini natural language system
- Based on Core Language Engine
- Full syntax and semantics for ATIS
- Coverage of whole corpus:
  - 70% syntax
  - 50% semantics

Using Natural Language Constraints

- Gemini natural language system
- Run pattern matcher
- For each sentence it returned
  - Remove fragment sentences
  - Leaving 179 repairs, 176 false positives
  - Parse each sentence
    - If succeed: mark as false positive
    - If fail:
      - run pattern matcher, make corrections
      - Parse again
      - If succeeds, mark as repair
      - If fails, mark no opinion

NL Constraints

- Syntax Only
  - Precision: 96%
- Syntax and Semantics
  - Correction: 70%

Recent work: EARS Metadata Evaluation (MDE)

- A recent multiyear DARPA bakeoff
- Sentence-like Unit (SU) detection:
  - find end points of SU
  - Detect subtype (question, statement, backchannel)
- Edit word detection:
  - Find all words in reparandum (words that will be removed)
- Filler word detection
  - Filled pauses (uh, um)
  - Discourse markers (you know, like, so)
  - Editing terms (I mean)
- Interruption point detection

Liu et al 2003

Kinds of disfluencies

- Repetitions
  - I * I like it
- Revisions
  - We * I like it
- Restarts (false starts)
  - It’s also * I like it

MDE transcription

- Conventions:
  - ./ for statement SU boundaries,
  - <> for fillers,
  - [] for edit words,
  - * for IP (interruption point) inside edits
- And <uh> <you know> wash your clothes wherever you are ./ and [you] * you really get used to the outdoors ./

MDE Labeled Corpora

	CTS	BN
Training set (words)	484K	182K
Test set (words)	35K	45K
STT WER (%)	14.9	11.7
SU %	13.6	8.1
Edit word %	7.4	1.8
Filler word %	6.8	1.8

MDE Algorithms

- Use both text and prosodic features
- At each interword boundary
  - Extract Prosodic features (pause length, durations, pitch contours, energy contours)
  - Use N-gram Language model
  - Combine via HMM, Maxent, CRF, or other classifier

State of the art: SU detection

- 2 stage
  - Decision tree plus N-gram LM to decide boundary
  - Second maxent classifier to decide subtype
- Current error rates:
  - Finding boundaries
    - 40-60% using ASR
    - 26-47% using transcripts

State of the art: Edit word detection

- Multi-stage model
    - HMM combining LM and decision tree finds IP
    - Heuristics rules find onset of reparandum
    - Separate repetition detector for repeated words
- One-stage model
    - CRF jointly finds edit region and IP
    - BIO tagging (each word has tag whether is beginning of edit, inside edit, outside edit)
- Error rates:
    - 43-50% using transcripts
    - 80-90% using ASR

Using only lexical cues

- 3-way classification for each word
  - Edit, filler, fluent
- Using TBL
  - Templates: Change
    - Word X from L1 to L2
    - Word sequence X Y to L1
    - Left side of simple repeat to L1
    - Word with POS X from L1 to L2 if followed by word with POS Y

Rules learned

- Label all fluent filled pauses as fillers
- Label the left side of a simple repeat as an edit
- Label “you know” as fillers
- Label fluent well’s as filler
- Label fluent fragments as edits
- Label “I mean” as a filler

Error rates using only lexical cues

- CTS, using transcripts
  - Edits: 68%
  - Fillers: 18.1%
- Broadcast News, using transcripts
  - Edits 45%
  - Fillers 6.5%
- Using speech:
  - Broadcast news filler detection from 6.5% error to 57.2%
- Other systems (using prosody) better on CTS, not on Broadcast News

Conclusions: Lexical Cues Only

- Can do pretty well with only words
  - (As long as the words are correct)
- Much harder to do fillers and fragments from ASR output, since recognition of these is bad

Fragments

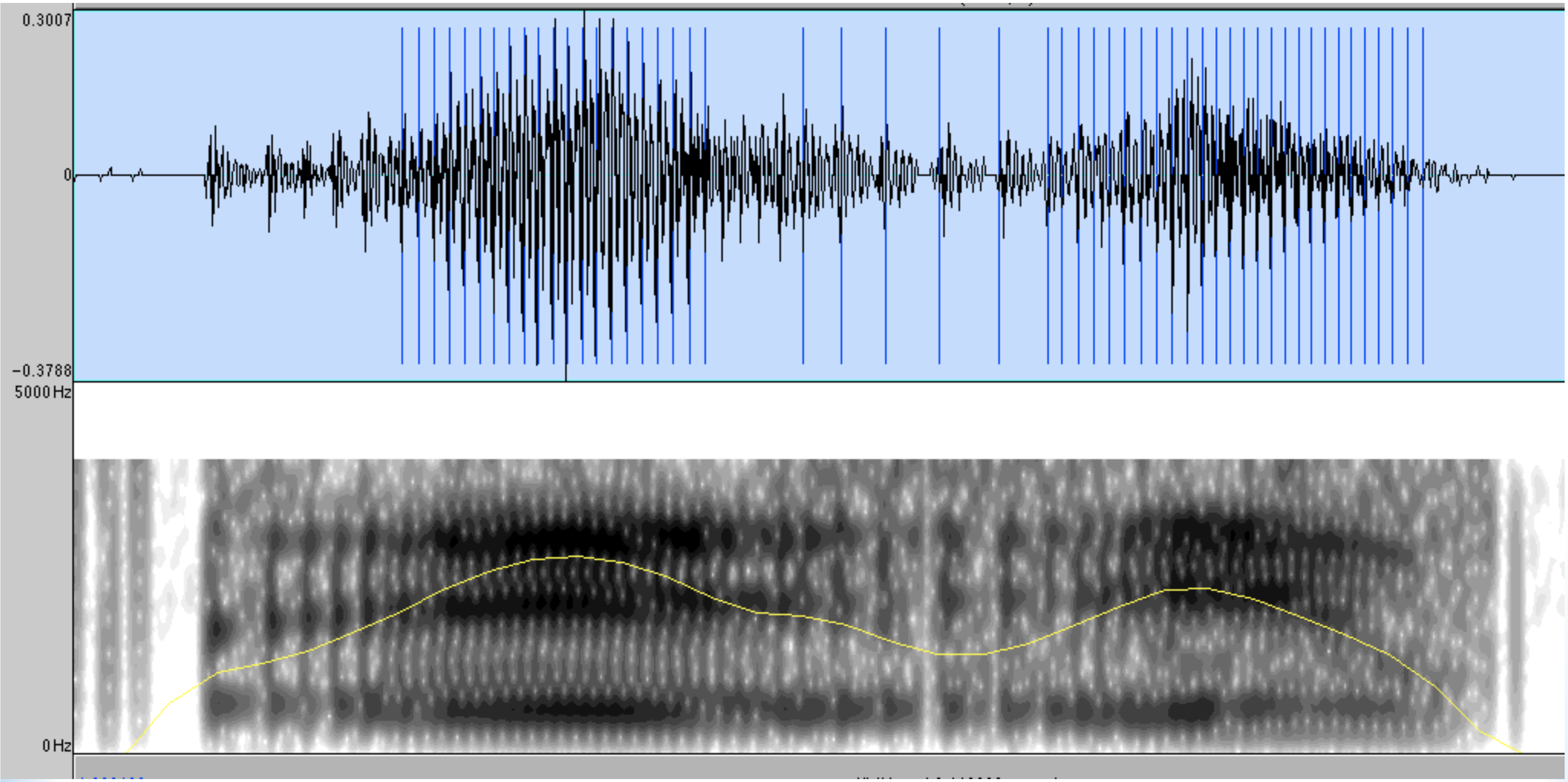
- Incomplete or cut-off words:
  - Leaving at seven fif- eight thirty
  - uh, I, I d-, don't feel comfortable
  - You know the fam-, well, the families
  - I need to know, uh, how- how do you feel…

Uh yeah, yeah, well, it- it- that's right. And it-

- SWBD: around 0.7% of words are fragments (Liu 2003)
- ATIS: 60.2% of repairs contain fragments (6% of corpus sentences had a least 1 repair) Bear et al (1992)
- Another ATIS corpus: 74% of all reparanda end in word fragments (Nakatani and Hirschberg 1994)

Fragment glottalization

- Uh yeah, yeah, well, **it- it-** that's right. And it-



Why fragments are important

- Frequent enough to be a problem:
  - Only 1% of words/3% of sentences
  - But if miss fragment, tend to get surrounding words wrong (word segmentation error).
  - Goldwater et al.:
    - 14% absolute increase in word error rate (from 18% to 32%) for words before fragments!!
- Useful for finding other repairs
  - In 40% of SRI-ATIS sentences containing fragments, fragment occurred at right edge of long repair
  - 74% of ATT-ATIS reparanda ended in fragments
- Sometimes are the only cue to repair
  - "leaving at <seven> <fif-> eight thirty"

How fragments are dealt with in current ASR systems

- In training, throw out any sentences with fragments
- In test, get them wrong
- Probably get neighboring words wrong too!
- !!!!!

Cues for fragment detection

- 49/50 cases examined ended in silence >60msec; average 282ms (Bear et al)
- 24 of 25 vowel-final fragments glottalized (Bear et al)
  - Glottalization: increased time between glottal pulses
- 75% don't even finish the vowel in first syllable (i.e., speaker stopped after first consonant) (O'Shaughnessy)

Cues for fragment detection

- Nakatani and Hirschberg (1994)
- Word fragments tend to be content words:

Lexical Class	Token	Percent
Content	121	42%
Function	12	4%
Untranscribed	155	54%

Cues for fragment detection

- Nakatani and Hirschberg (1994)
- 91% are one syllable or less

Syllables	Tokens	Percent
0	113	39%
1	149	52%
2	25	9%
3	1	0.3%

Cues for fragment detection

- Nakatani and Hirschberg (1994)
- Fricative-initial common; not vowel-initial

Class	% words	% frags	% 1-C frags
Stop	23%	23%	11%
Vowel	25%	13%	0%
Fric	33%	45%	73%

Liu (2003): Acoustic-Prosodic detection of fragments

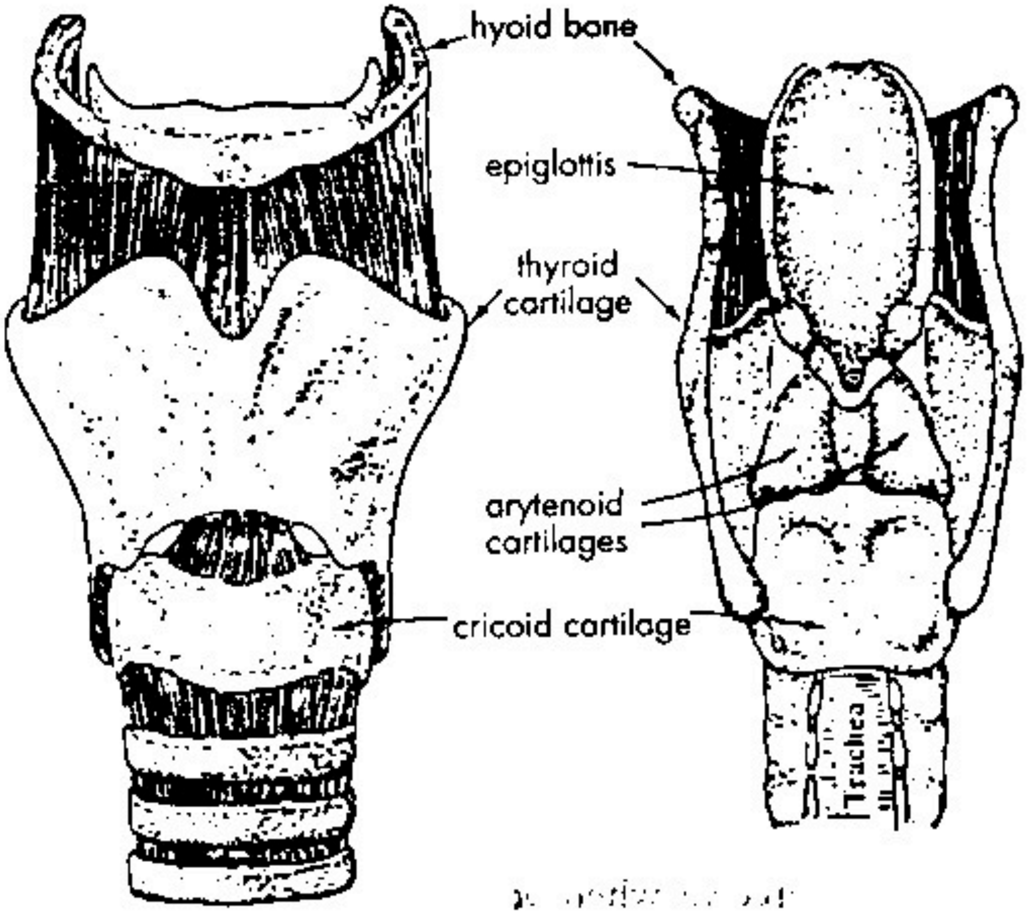
- Prosodic features
  - Duration (from alignments)
    - Of word, pause, last-rhyme-in word
    - Normalized in various ways
  - F0 (from pitch tracker)
    - Modified to compute stylized speaker-specific contours
  - Energy
    - Frame-level, modified in various ways

Liu (2003): Acoustic-Prosodic detection of fragments

- Voice Quality Features
  - Jitter
    - A measure of perturbation in pitch period
      - Praat computes this
  - Spectral tilt
    - Overall slope of spectrum
    - Speakers modify this when they stress a word
  - Open Quotient
    - Ratio of times in which vocal folds are open to total length of glottal cycle
    - Can be estimated from first and second harmonics
    - Creaky voice (laryngealization) vocal folds held together, so short open quotient

The larynx

main function of vocal folds: block objects from falling into trachea

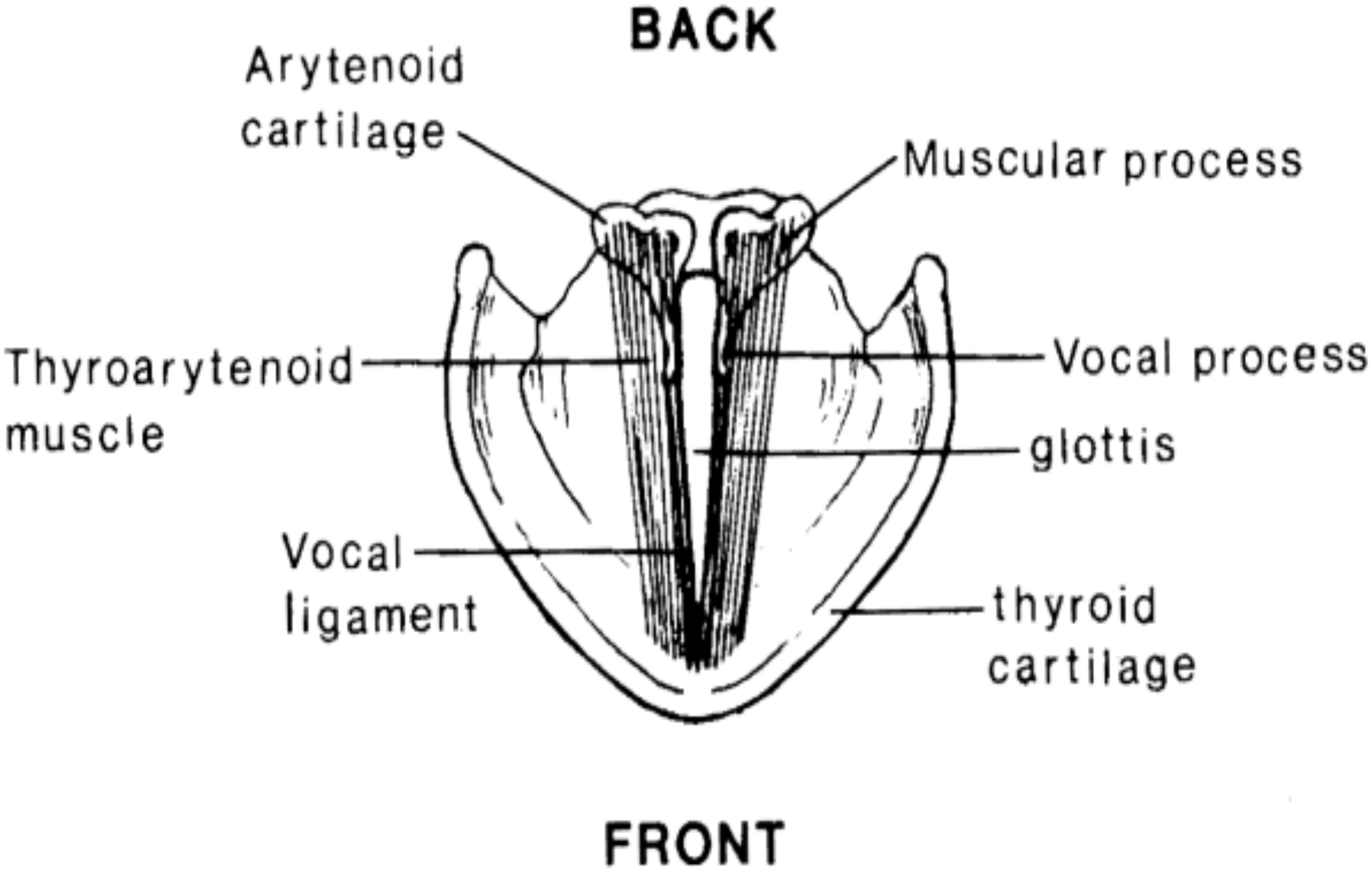


[illegible]

[illegible]

Slide from Ulrike Gut

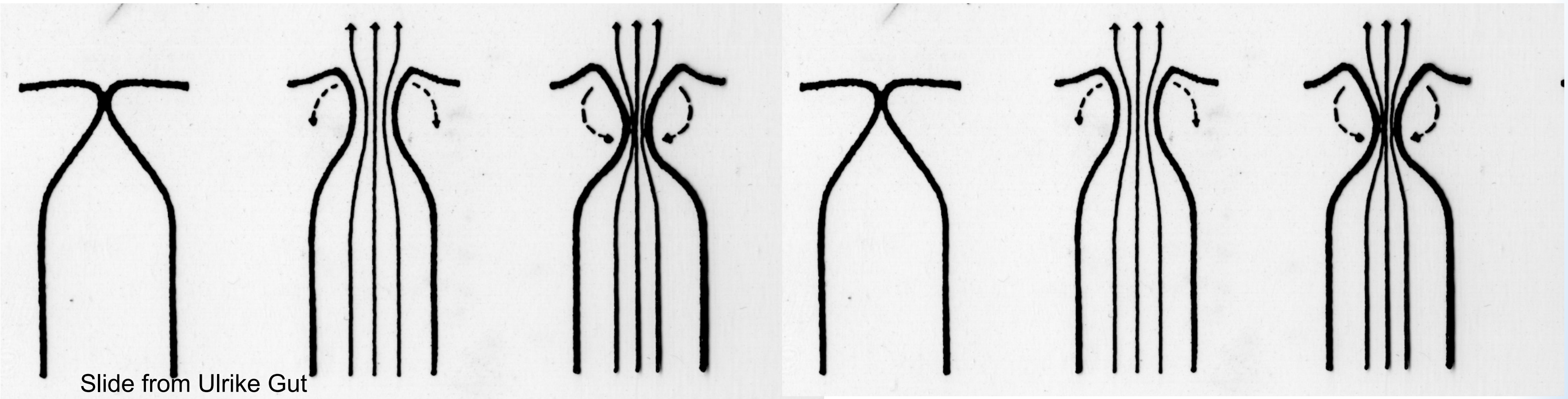
Inside the larynx



Slide from Ulrike Gut

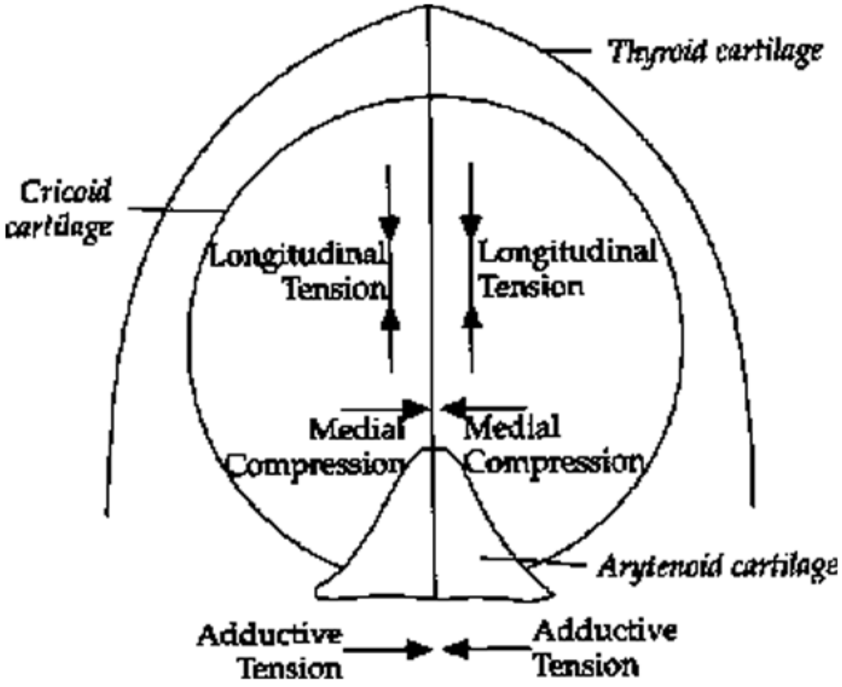
Phonation

- phonation: vibration (=opening and closing) of the vocal folds
  - vocal folds closed - air from the lungs pushes them apart – sucked back together (Bernoulli effect)



Slide from Ulrike Gut

Voice Quality and the Larynx



Adductive tension

(interarytenoid muscles adduct the arytenoid muscles)

Medial compression

(adductive force on vocal processes- adjustment of ligamental glottis)

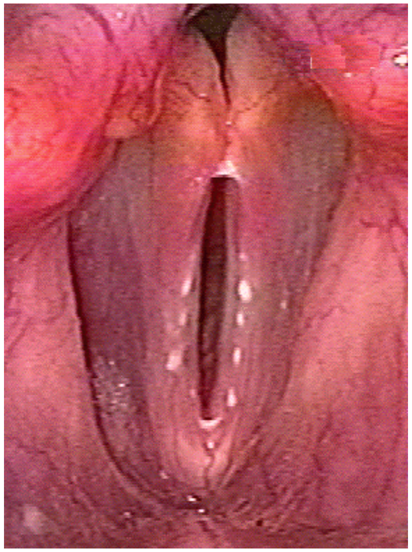
Longitudinal pressure

(tension of vocal folds)

Slide from K. Marasek, J. Wilcox

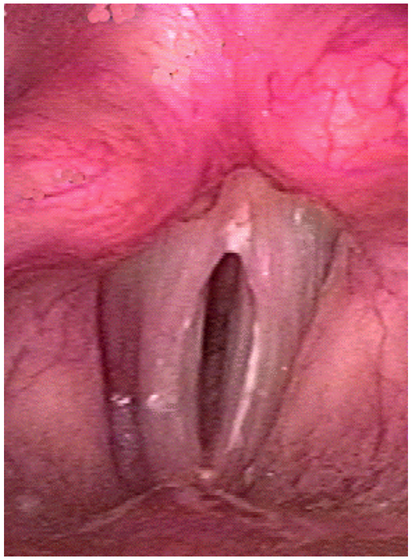
Modulation of vocal fold vibration

- vocal folds are moved (adducted) by muscles
- can be tensed – the shorter the vocal folds the faster they vibrate



200 times/sec

120 times/sec



Slide from Ulrike Gut

Modes of phonation

- voicelessness = no vocal fold vibration
- modal (normal) voicing
- whisper
- breathy voice
- voice
- creaky voice

Slide from Ulrike Gut

Modal voice

neutral mode

muscular adjustments are moderate

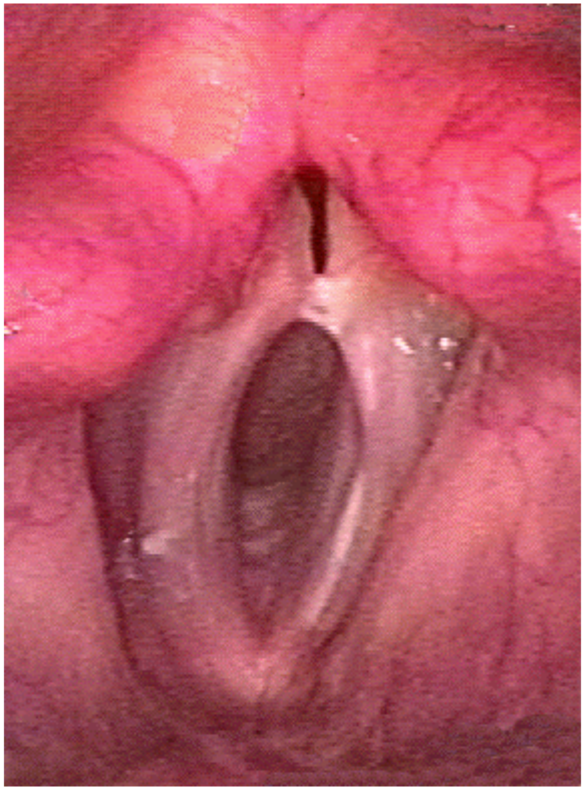
vibration of the vocal folds is periodic with full closing of glottis, so no audible friction noises are produced when air flows through the glottis.

frequency of vibration and loudness are in the lowto mid range for conversational speech

Slide from K. Marasek, J. Wilcox

Breathy voice

- arytenoid cartilages remain slightly apart
- continuous airflow during vocal fold vibration



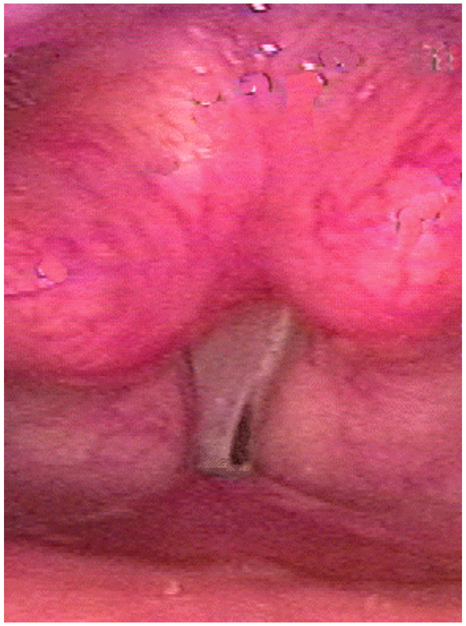
Slide from Ulrike Gut

Creaky voice

- arytenoid cartilages tightly together so that vocal folds can only vibrate at the other end

normal

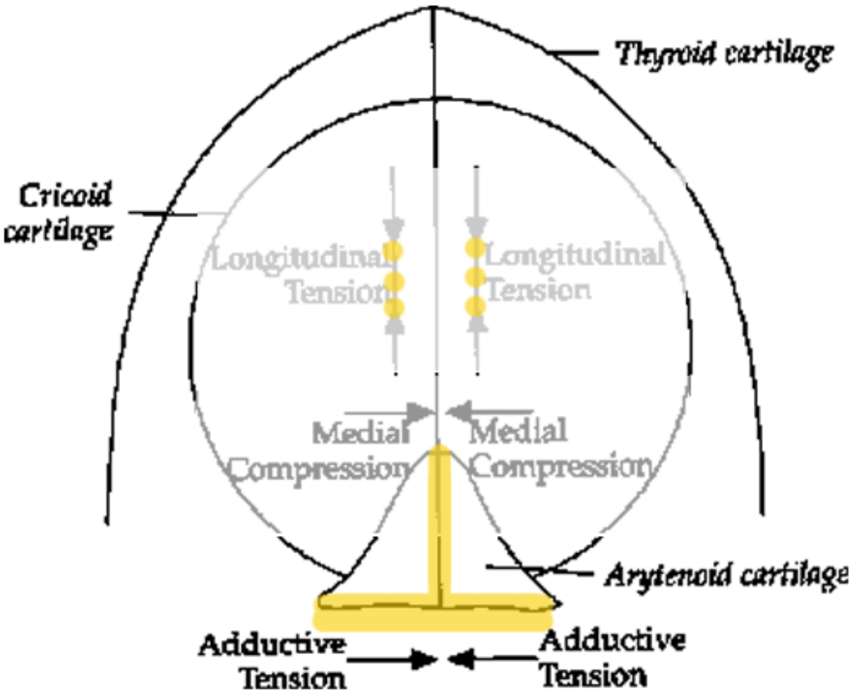
creaky voice



Slide from Ulrike Gut

Creaky voice – voiced phonation

- vocal folds vibrate at a very low frequency – vibration is somewhat irregular, vibrating mass is “heavier” because of low tension (only the ligamental part of glottis vibrates)
- The vocal folds are strongly adducted
- longitudinal tension is weak
- Moderately high medial compression
- Vocal folds “thicken” and create an unusually thick and slack structure.



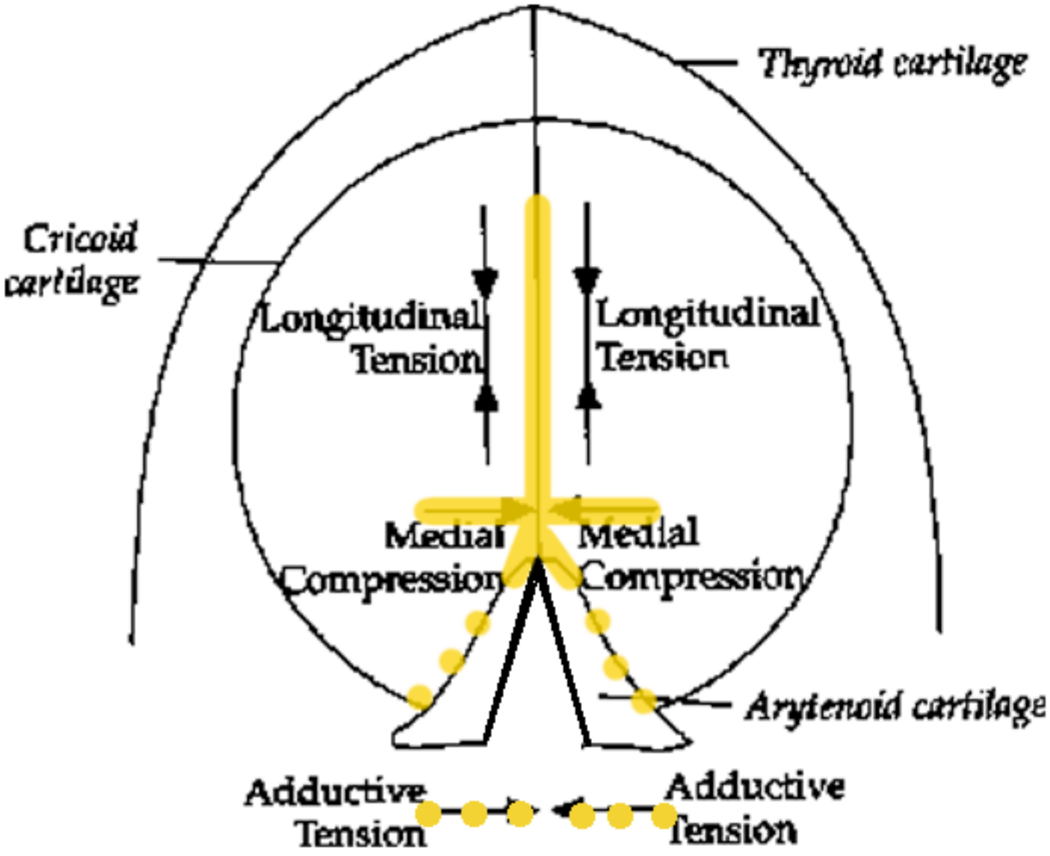
Slide from K. Marasek, J. Wilcox

Whisper

- in *whisper* there is no true vibration of the vocal folds; adduction of vocal folds while maintaining an opening between the arytenoid cartilages

Slide from Ulrike Gut

Whispery voice – voiceless phonation



Very low adductive tension

Medial compression moderately high

Longitudinal tension moderately high

Little or no vocal fold vibration

(produced through turbulences generated by the friction of the air in and above the larynx, which produces frication)

Liu (2003)

- Use Switchboard 80%/20%
- Downsampled to 50% frags, 50% words
- Generated forced alignments with gold transcripts
- Extract prosodic and voice quality features
- Train decision tree

Liu (2003) results

- Precision 74.3%, Recall 70.1%

		hypothesis	
		complete	fragment
reference	complete	109	35
	fragment	43	101

Liu (2003) features

- Features most queried by DT

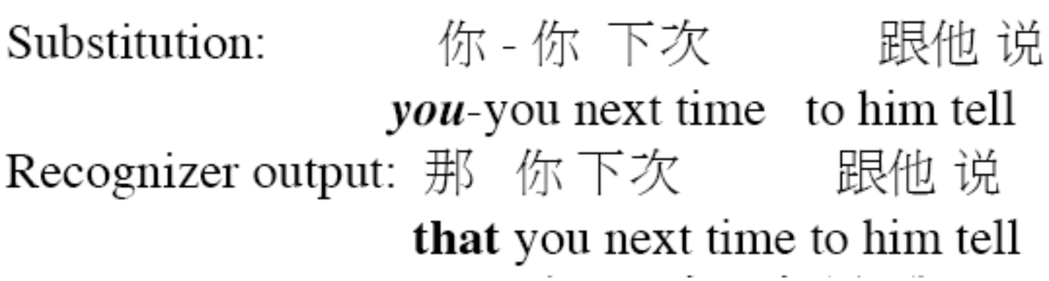
Feature	%
jitter	.272
Energy slope difference between current and following word	.241
Ratio between F0 before and after boundary	.238
Average OQ	.147
Position of current turn	0.084
Pause duration	0.018

Liu (2003) conclusion

- Very preliminary work
- Fragment detection is good problem that is understudied!

Fragments in other languages

- Mandarin (Chu, Sung, Zhao, Jurafsky 2006)
- Fragments cause similar errors as in English:



Substitution: 你 - 你 下次 跟他 说

you-you next time to him tell

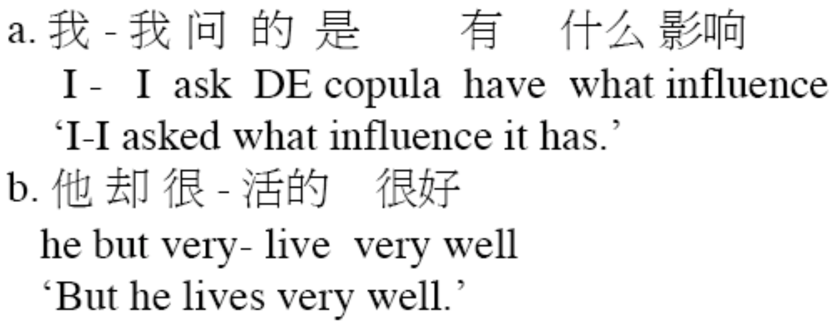
Recognizer output: 那 你 下次 跟他 说

that you next time to him tell

- 但我- 我問的是
- I- I was asking…
- 他是- 卻很- 活的很好
- He very- lived very well

Fragments in Mandarin

- Mandarin fragments unlike English; no glottalization.
- Instead: Mostly (unglottalized) repetitions



a. 我 - 我 问 的 是 有 什么 影响
I - I ask DE copula have what influence
'I-I asked what influence it has.'

b. 他 却 很 - 活的 很好
he but very- live very well
'But he lives very well.'

- So: best features are lexical, rather than voice quality

Summary

- Disfluencies
- Characteristics of disfluences
- Detecting disfluencies
- MDE bakeoff
- Fragments