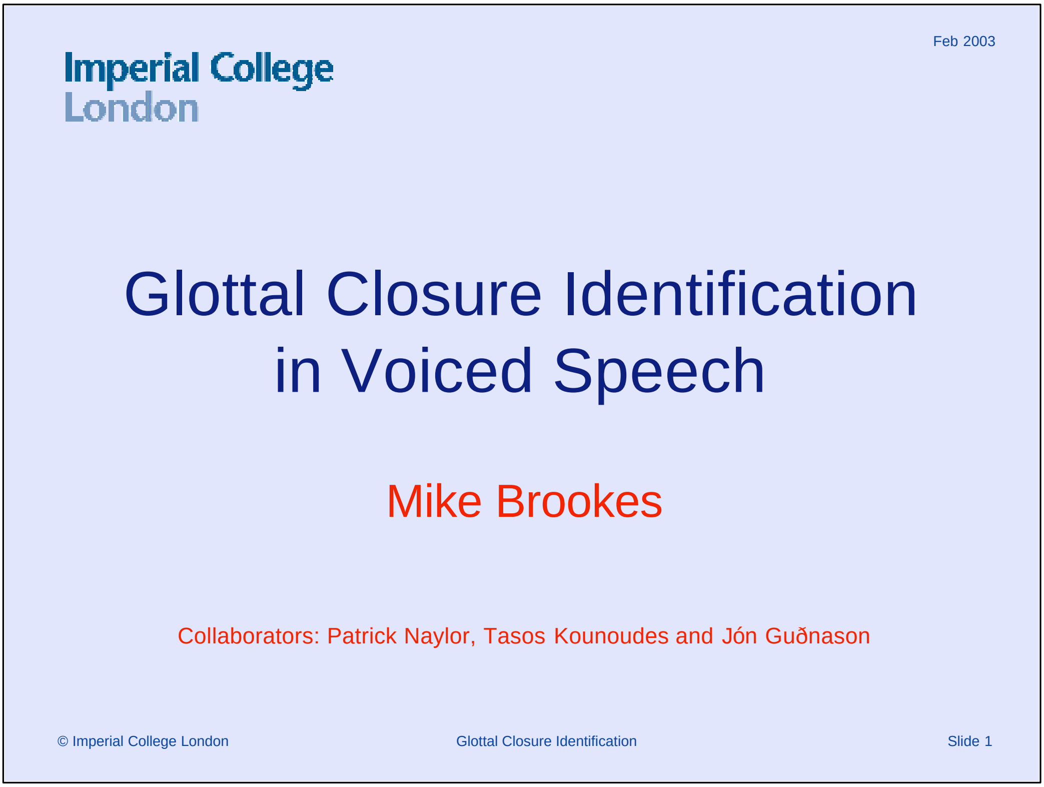


Feb 2003

Imperial College
London

Glottal Closure Identification in Voiced Speech

Mike Brookes

Collaborators: Patrick Naylor, Tasos Kounoudes and Jón Guðnason

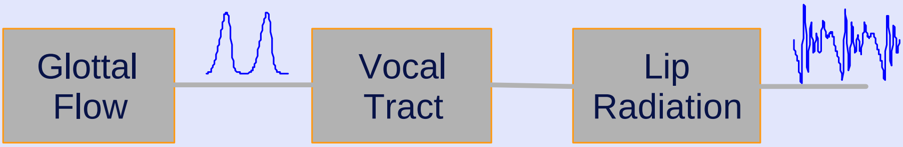
Abstract:

Glottal Closure Identification in Voiced Speech

Having the ability to identify the instants of glottal closure in voiced speech enables the use of larynx synchronous processing techniques such as closed-phase LPC analysis. These techniques make it possible to separate the characteristics of the glottal excitation waveform from those of the vocal tract filter and to treat the two independently in subsequent processing. Applications include low bit-rate coding, data-driven techniques for speech synthesis, prosody extraction, voice morphing, speaker normalization and speaker recognition.

This talk will describe a two-stage technique for determining the glottal closure instants from the speech waveform. In the first stage, candidate closure instants are identified from the group delay of the LPC residual; in the second stage dynamic programming is used to eliminate spurious candidates. The results obtained are compared with reference closure instants derived from the direct measurement of larynx activity.

Voiced Speech Production Model



- Why identify glottal closures ?
  - – Frame synchronisation in synthesis
  - – Larynx-synchronous processing
    - Speaker characterisation for synthesis
    - Low bit-rate speech coding
    - Speaker normalization and recognition

Why:

•Data driven techniques for synthesis

•Closed-phase analysis of speech – separate excitation from tract

•speaker recognition

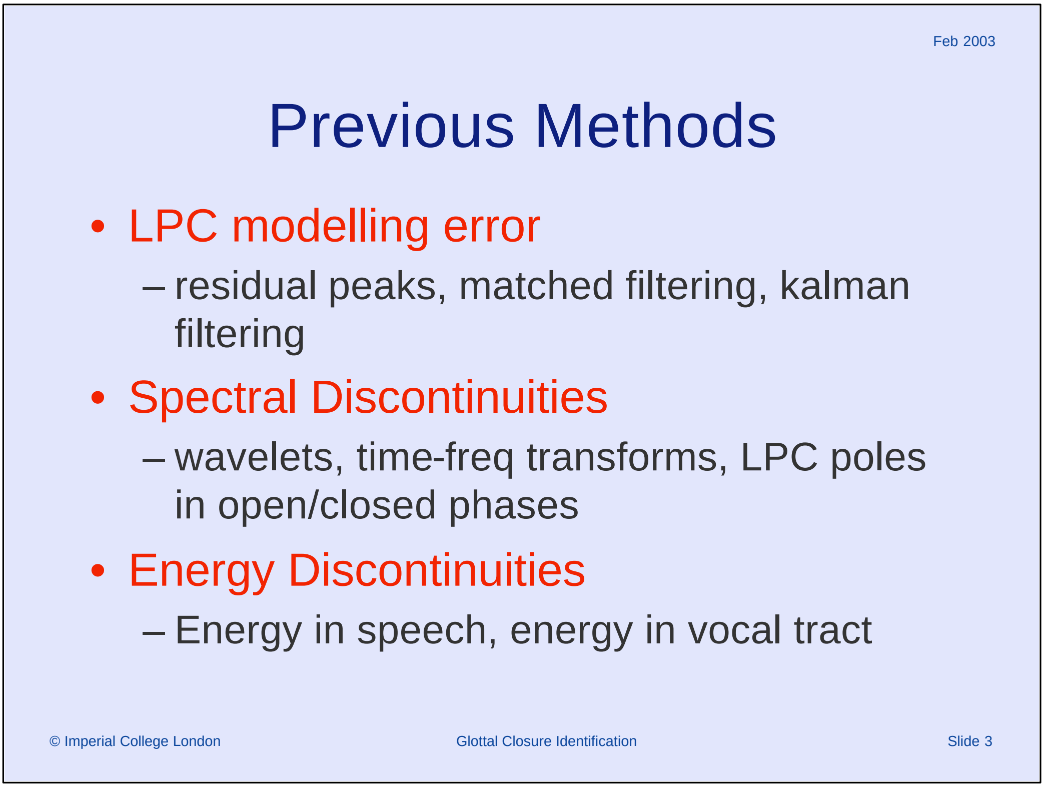
•speaker normalisation

•voice morphing

•prosody extraction

•voice source analysis

•low bit-rate coding



Previous Methods

- LPC modelling error
  – residual peaks, matched filtering, kalman filtering
- Spectral Discontinuities
  – wavelets, time-freq transforms, LPC poles in open/closed phases
- Energy Discontinuities
  – Energy in speech, energy in vocal tract

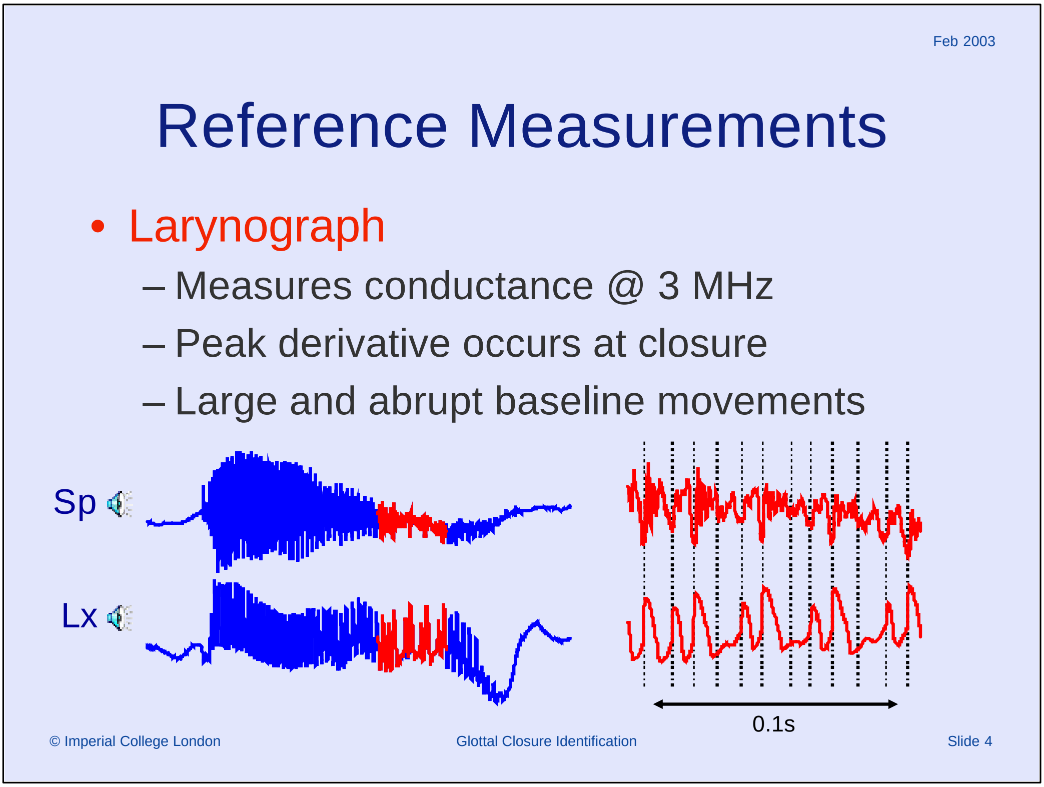
•In general, all methods work by transforming the speech signal into some alternative domain and then trying to detect the discontinuities that occur at glottal closure.

•LPC methods relay on the fact that an AR model is a good representation for vowels and not bad for voiced consonants.

•The time-frequency methods are trying to automate recognition of the characteristic of spectrograms which show a stong wide-bandwidth energy pulse at closure.

•Distinction between published results the estimate specific instants and those that generate a waveform + comment that “you can clearly see that this identifies the excitation instants”.

•Many (most) speech segments are quite easy to process – the difficulty is to make a robust method that works with the difficult ones. It s easy to devise a technique that works well on synthetic speech.



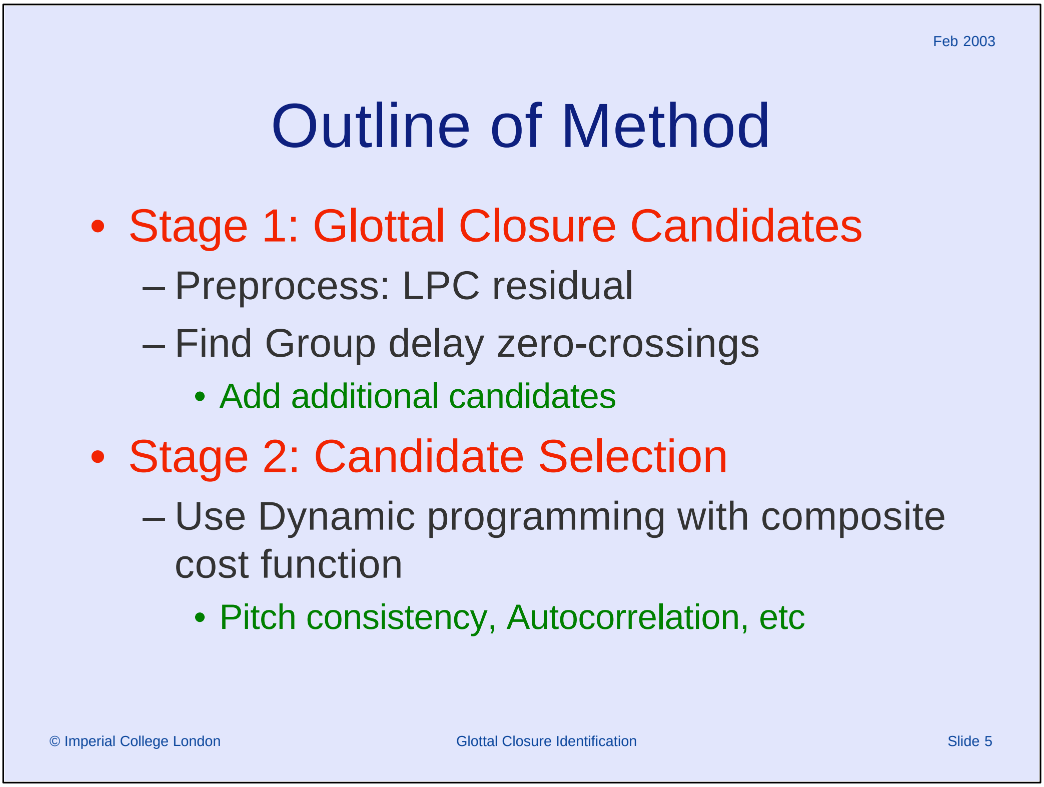
Reference Measurements

- Larynograph
  - Measures conductance @ 3 MHz
  - Peak derivative occurs at closure
  - Large and abrupt baseline movements

In the detail section:

- first few cycles all going well, larynx closure marked by an abrupt increase in conductance. If you magnify it, you will find that there is a 1 ms delay in the speech signal corresponding to the distance from larynx to microphone.
- Closure = peak derivative, opening = falls to 30% of peak conductance increase. There may be a secondary excitation at opening.
- Several cycles have incomplete closures as the speech intensity falls
- Identifying closure instants from the Lx waveform is not trivial:

1. baseline moves around primarily due to gross movement of the larynx – can’t just low-pass filter because changes can be abrupt.
2. amplitude varies greatly particularly at voice offset – can’t just use a threshold
3. larynx frequency varies by more than a factor of 10: 40 Hz to 500 Hz.

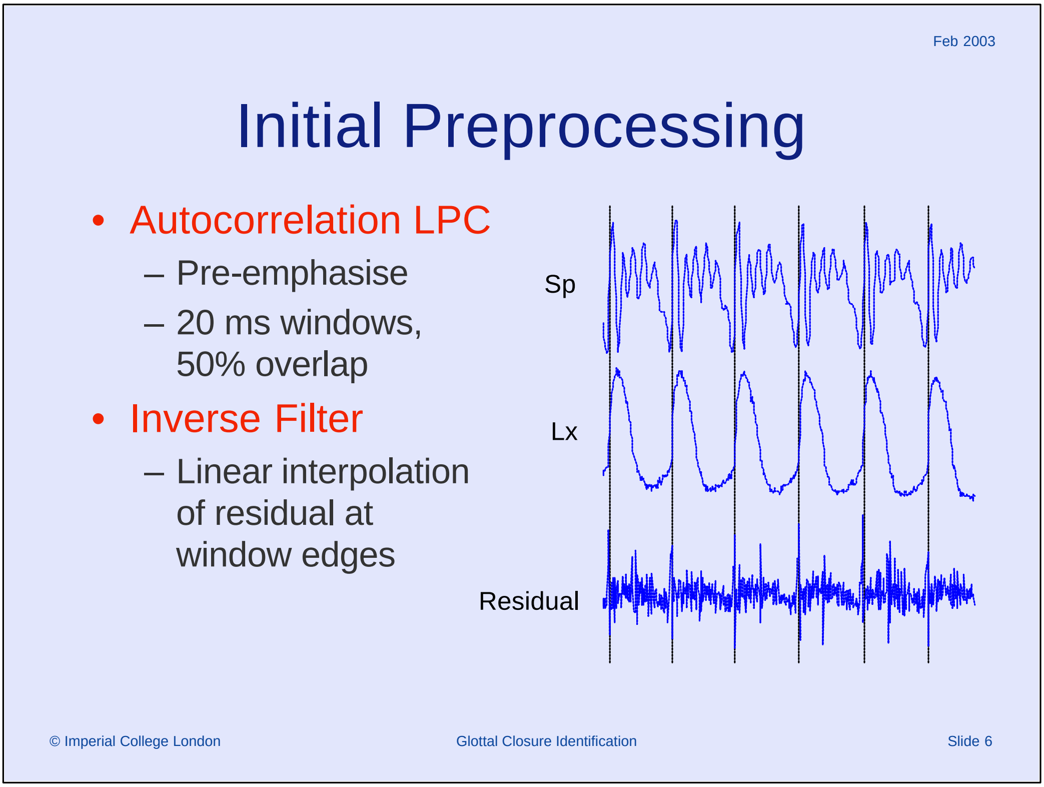


Outline of Method

- Stage 1: Glottal Closure Candidates
  - Preprocess: LPC residual
  - Find Group delay zero-crossings
    - Add additional candidates
- Stage 2: Candidate Selection
  - Use Dynamic programming with composite cost function
    - Pitch consistency, Autocorrelation, etc

• In the first stage we determine a set of candidate glottal closure instants. We want to be inclusive at this stage since any missed closures will never be recovered. We base our detection on the LPC residual and look at the variation of the average group delay in a sliding window – of which more later.

• In the second stage, we use DP to select a subsequence of the candidates identified in stage 1 as our final outputs.



Initial Preprocessing

- Autocorrelation LPC
  - – Pre-emphasise
  - – 20 ms windows, 50% overlap
- Inverse Filter
  - – Linear interpolation of residual at window edges

Sp
Lx
Residual

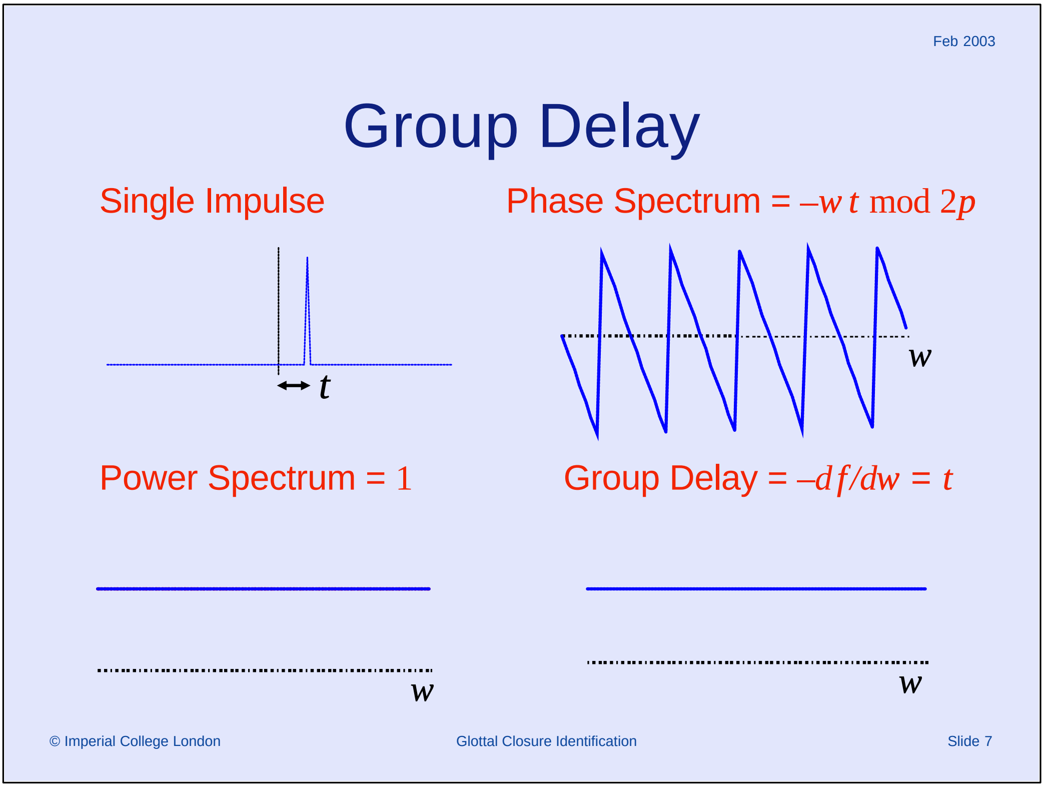
•This slide the Speech, Lx and LPC residual waveforms for a well-behaved vowel segment .

•The Lx has been delayed by 1 ms to compensate for the larynx-to-microphone delay.

•The residual is approximately the 2nd derivative of the glottal airflow waveform.

•The residual shows a pronounced spike at each glottal closure. Quite a lot of noise and sometimes an additional spike at glottal opening.

•We want to detect the impulses that occur in the residual at glottal closures. There are others that occur at opening and sometimes at other times.



Group Delay

Single Impulse

Phase Spectrum = $-w\,t \bmod 2p$

Power Spectrum = 1

Group Delay = $-df/dw$ = t

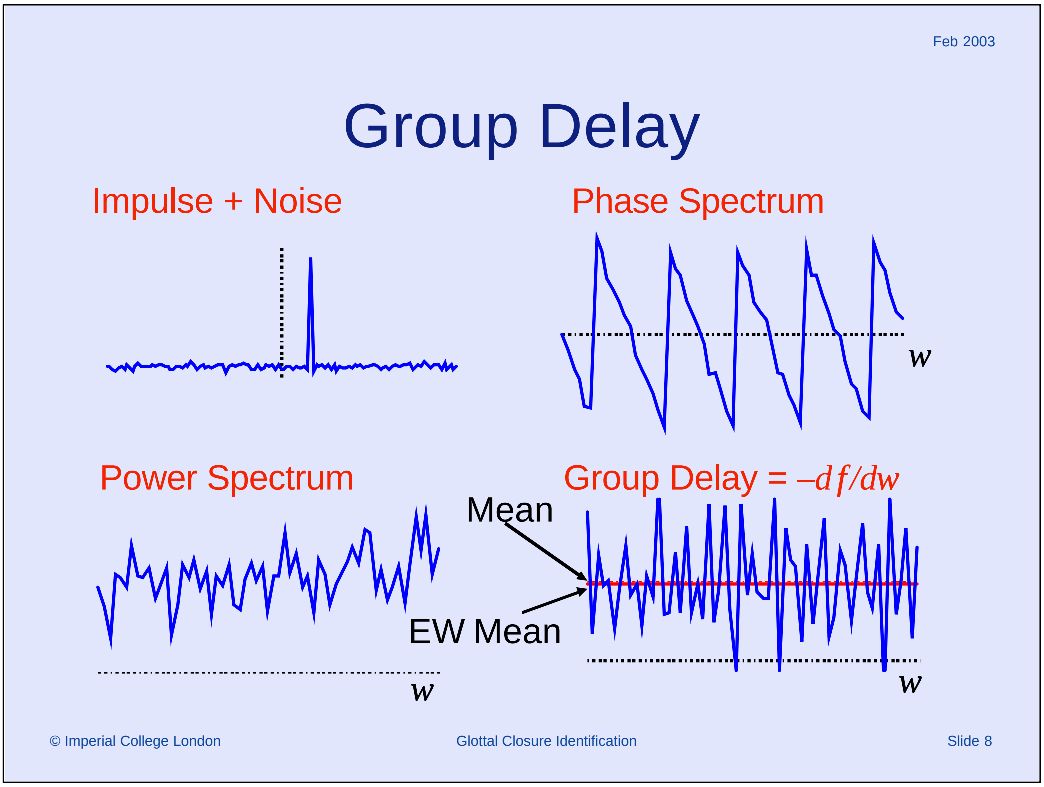
•The idea of using the group delay to locate impulses in the LPC residual was first proposed by Smits & Yegnanarayana (1995) and independently by Stylianou at Bell Labs in 1999.

•This slide illustrates the idea with an input signal consisting of a single impulse that is a time τ after the time origin at the centre of the window.

•The power spectrum of this signal is flat and the phase spectrum is equal to $-wt \bmod 2p$.

•If we calculate the group delay as $-df/dw$ we find that this has a constant value of t at all frequencies.

Group Delay



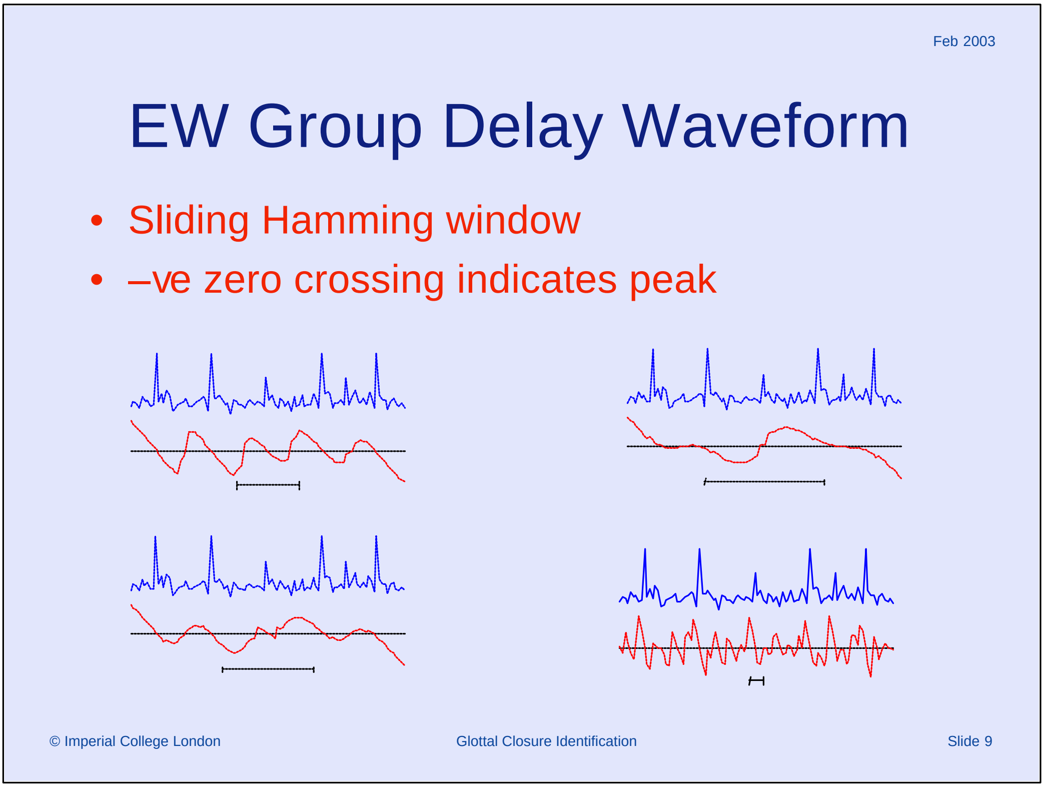
Impulse + Noise

Phase Spectrum

Power Spectrum

Group Delay = $-df/dw$

- Sadly, even a small amount of noise completely destroys the group delay function, so we average it over frequency to get a good estimate.
- There are some interesting relationships between the group delay and the time domain:
    - The group delay at w=0 is equal to the position of the centre of gravity of the input signal.
    - The energy weighted average group delay, that is, weighted by the power spectrum, is equal to the position of the centre of gravity of the input signal squared.
- Smits uses mean group delay
- Stylianou uses group delay at w=0 – not very robust
- We use energy weighted mean: robust and can be calculated in the time domain: just the centre of gravity of the input signal.



EW Group Delay Waveform

- Sliding Hamming window
- –ve zero crossing indicates peak

•Impulses plus noise

•Top left: window width equls period

•get a nice zero crossing for each impulse

•If noise-free impulse, then slope will be unity

• If noise present, then slope will be less than unity

• Bottom left: window =~1.5 period

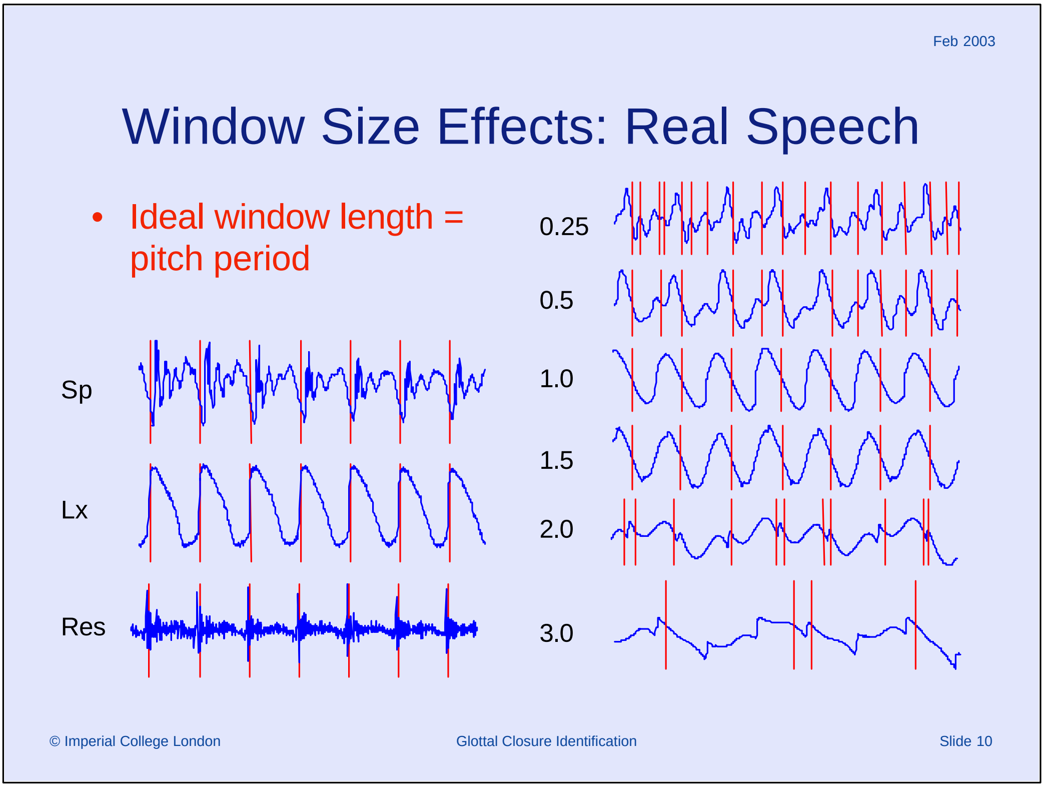
•still OK but small impulse is much shorter + spurious impulse pulls neighbours towards it

•Top right: window = 2*period: several zero-crossings have disappeared

•Bottom right: window very short: many additional zero crossings, slopes are all near -1.

****

Graph scales: V50%, H80%



Window Size Effects: Real Speech

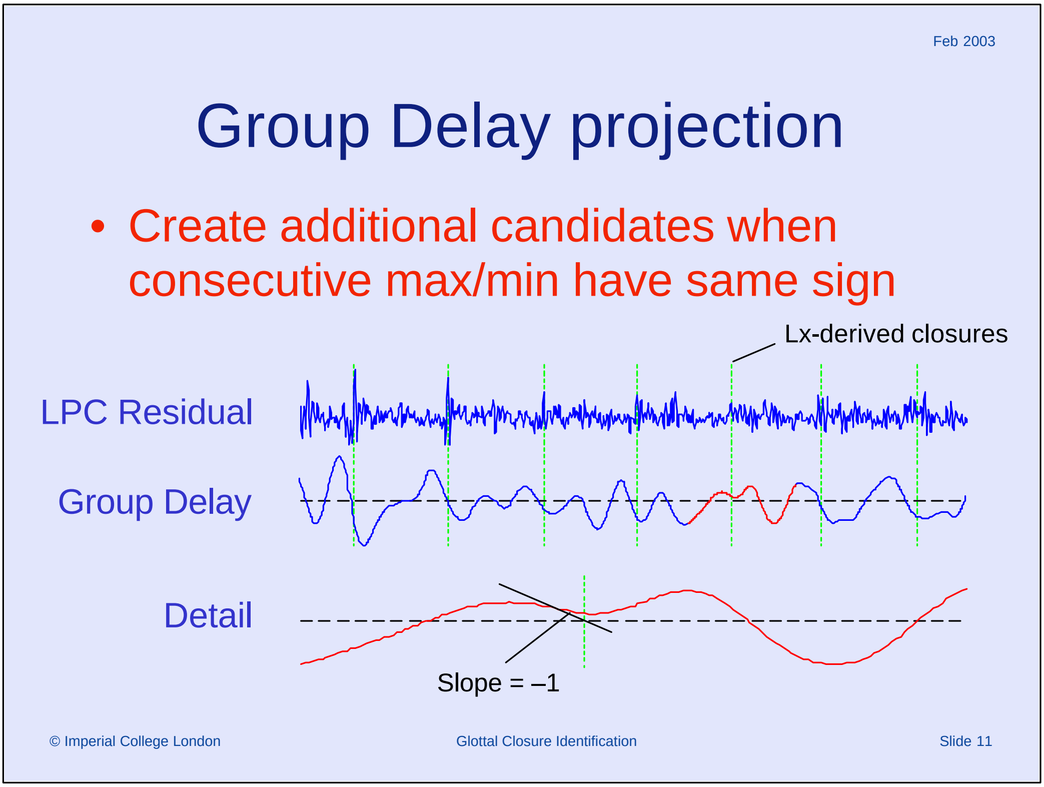
- Ideal window length = pitch period

• On the left, the slide shows a real speech signal with the corresponding Lx and LPC residual waveforms.

•On the right, we show the energy-weighted group delay waveform for different window lengths, normalized to the larynx period:

> 1 pitch period – everything is wondwerful with a single well-defined zero-crossing per larynx cycle. The vertical bars indicate the zero-crossings and these have been copied onto the Speech and Lx waveforms for comparison.
>
> 1.5 pitch periods – still pretty good.
>
> 2 pitch periods. The problem here is that when one closure is in the centre of the sliding window, one is just entering on the right and another is just leaving on the left. These cause a jump in the group delay waveform which gives rise to multiple zero-crossings for some cycles.
>
> 3 pitch periods – all gone to hell.
>
> <1 pitch period results in an increasing number of false alarms but we still include the correct instants.

Moral: window must not be longer than two pitch periods. We use 6 ms which therefore works up to 333 Hz larynx frequency.



Group Delay projection

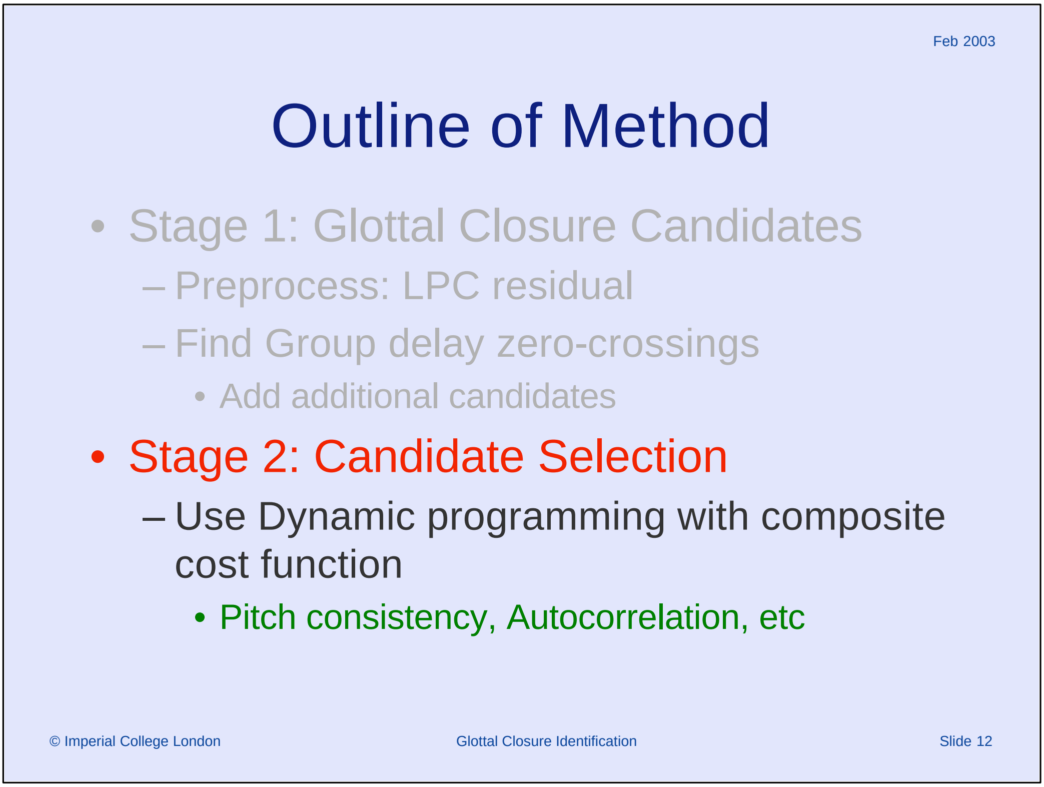
- Create additional candidates when consecutive max/min have same sign

•We have found that, if you get the window size right, you do pick up almost all the closures.

•Occasionally, however, you get a waveform like this in which a small excitation at the true closure is followed by a similar sized excitation at opening which prevents a zero-crossing.

•We therefore look at the group delay waveform and if it contains a maximum followed by a minimum without an intervening zero-crossing, we project the point mid-way between these two onto the time axis with a slope of -1 to create an additional closure candidate.

•In this way, as we shall see when I show the results, we pick up virually all the true closures.

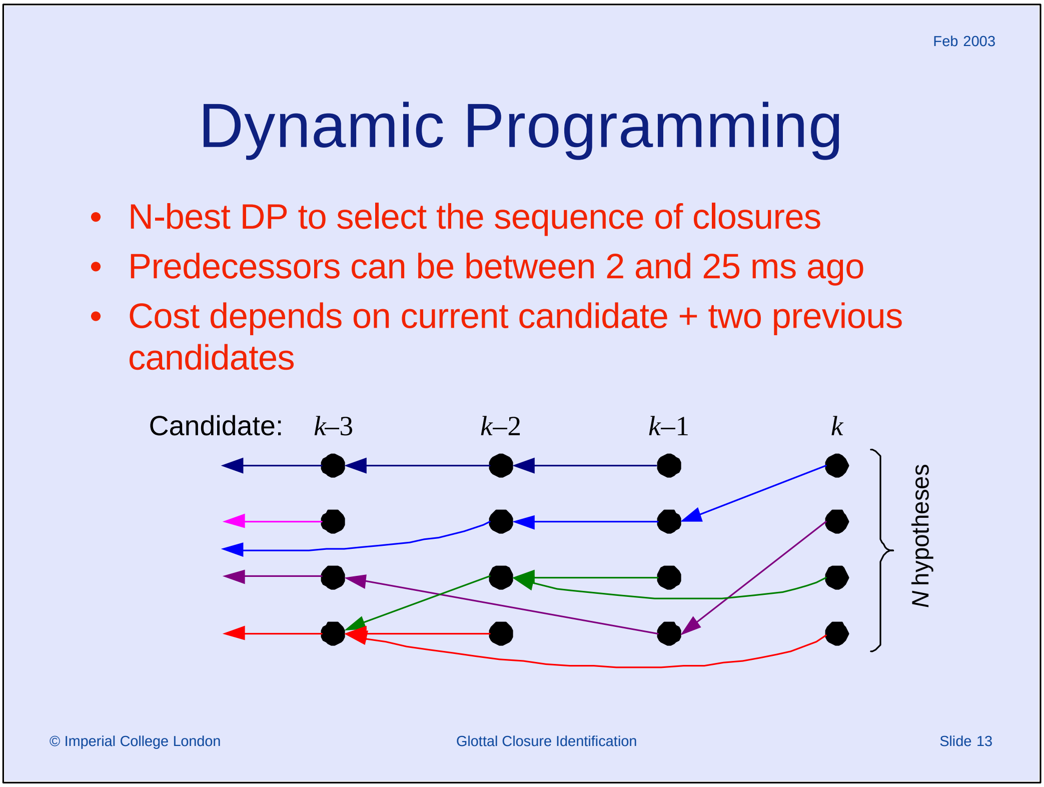


Outline of Method

- Stage 1: Glottal Closure Candidates
  - Preprocess: LPC residual
  - Find Group delay zero-crossings
    - Add additional candidates
- Stage 2: Candidate Selection
  - Use Dynamic programming with composite cost function
    - Pitch consistency, Autocorrelation, etc

•In stage 1, we have identified a sequence of candidate glottal closure instants. We hope there are none missing but accept that we have included some spurious ones.

•In stage 2 of the algorithm, we apply dynamic programming to select a subsequence that corresponds to the actual closures.



Dynamic Programming

- N-best DP to select the sequence of closures
- Predecessors can be between 2 and 25 ms ago
- Cost depends on current candidate + two previous candidates

• Assuming that candidate k is a genuine closure, we keep the N-best hypotheses about the sequence of previous closures.

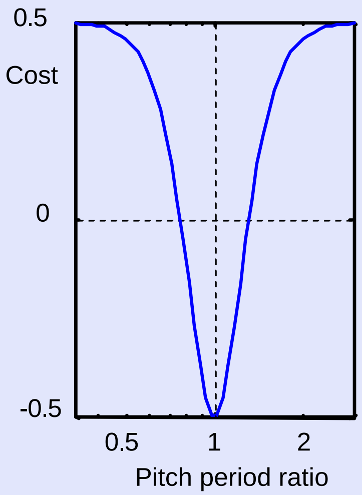
•Within a voiced speech segment, the closure that preceeds candidate k must occur between 2 and 25 ms before it. This corresponds to larynx frequencies between 40 and 500 Hz.

•The terms in the cost function depend on the current candidate position and the two previous candidates.

•There is no penalty for skipping over candidates. In order to prevent the minimum cost path just occupying as few closures as possible, the cost terms in the DP can be negative as well as positive.

Cost 1: Pitch Consistency

- Cost depends on the ratio of consecutive pitch periods
- Negative cost if ratio in range 0.81 to 1.24

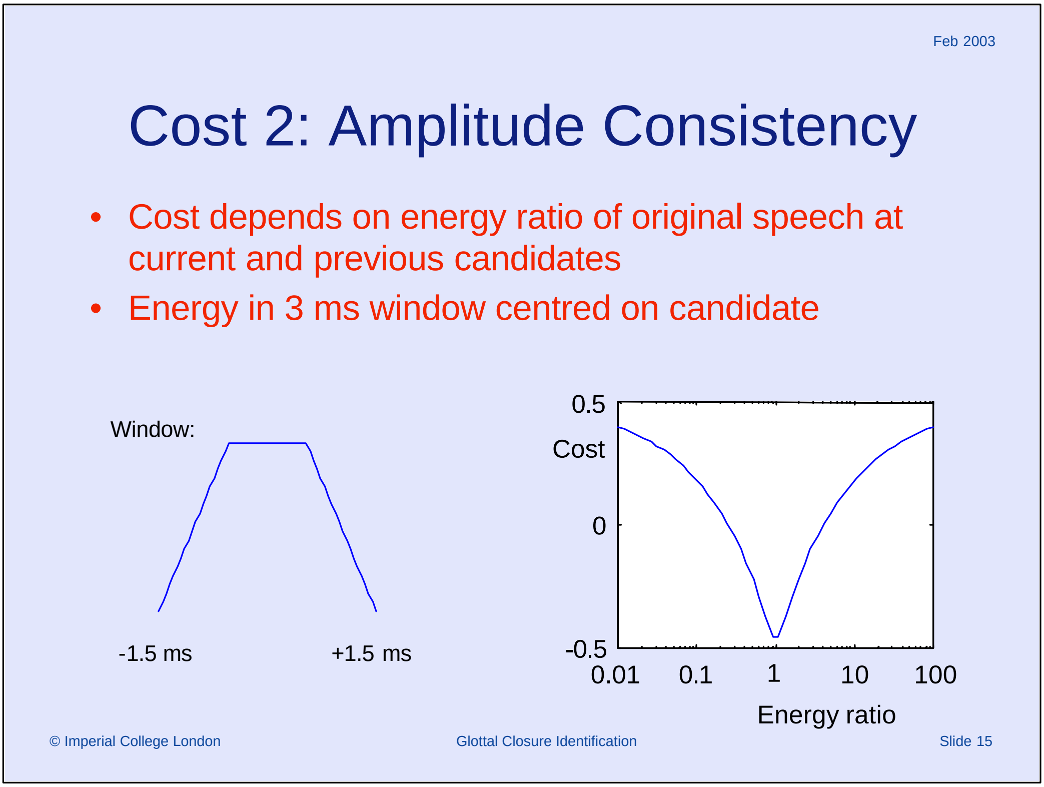


•This is the most important cost term.

•Look at three consecutive closures and take the ratio of the two time intervals.

•The lowest cost is if the ratio is unity.

•We probably should also have a dip in the graph at 0.5 and 2 to account for the pitch halving that occurs in creaky voice.



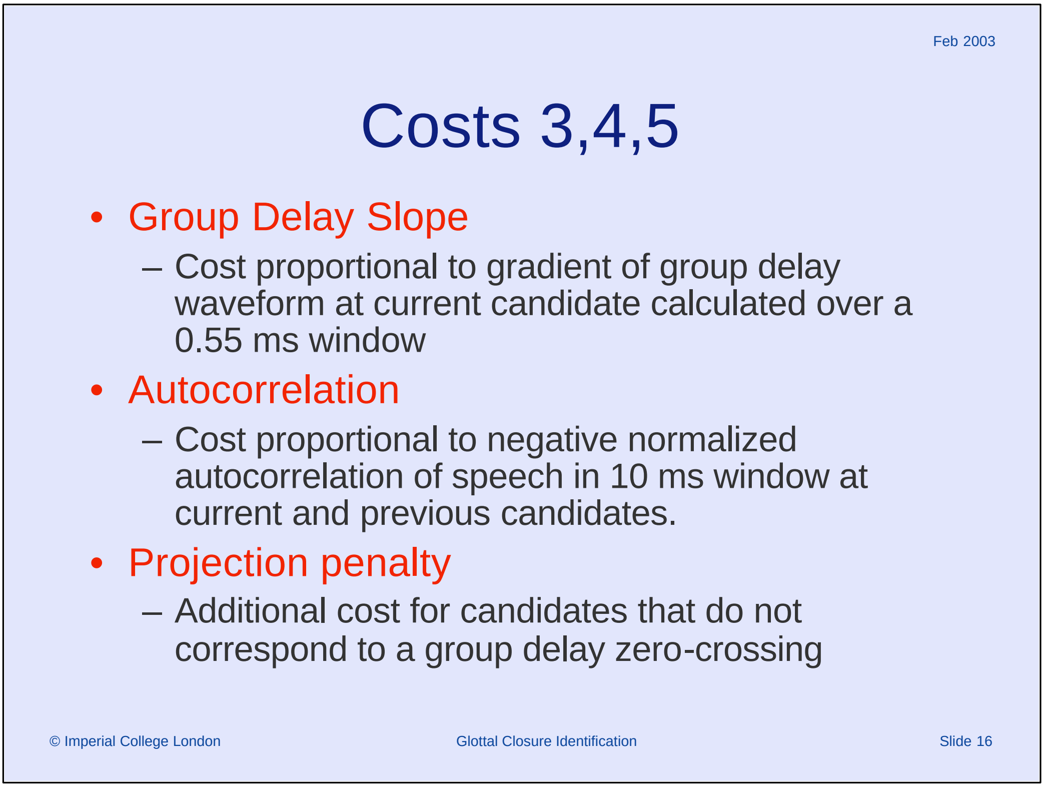
Cost 2: Amplitude Consistency

- Cost depends on energy ratio of original speech at current and previous candidates
- Energy in 3 ms window centred on candidate

Window:
-1.5 ms
+1.5 ms
0.5
Cost
0
-0.5
0.01
0.1
1
10
100
Energy ratio

Look at the energy of the original speech in a narrow window centred on the current and previous candidates.

We get a negative cost if the amplitude is within a factor of 3 or so



Costs 3,4,5

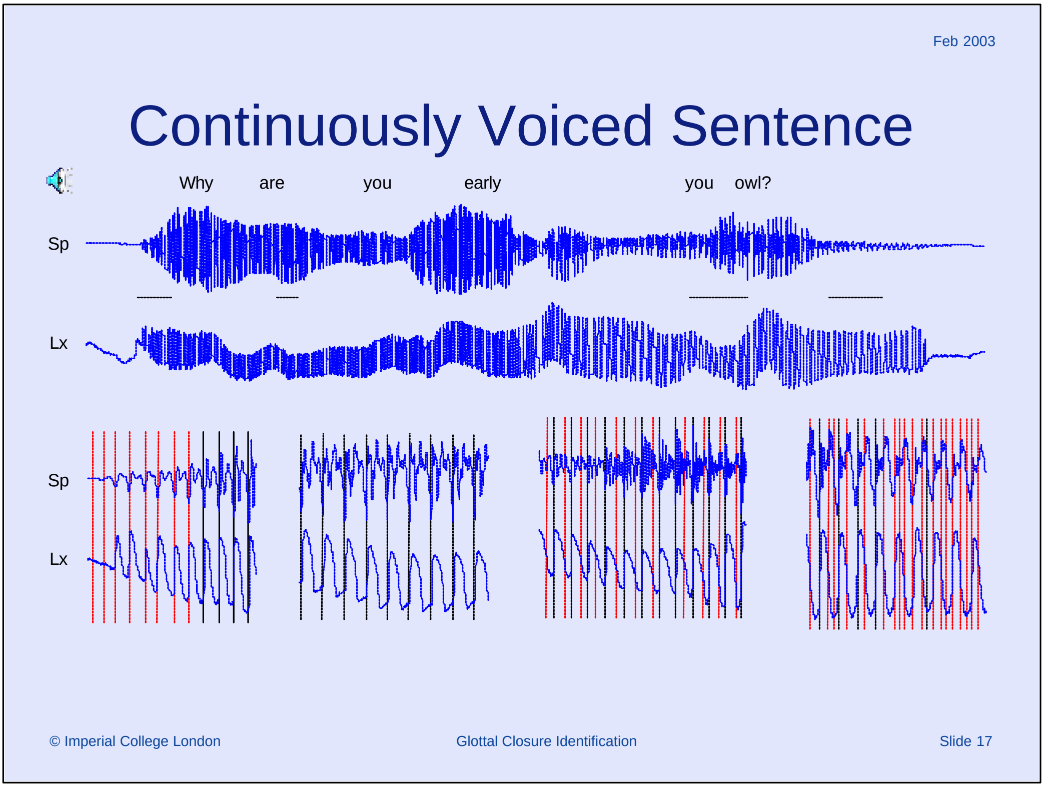
- Group Delay Slope
  - Cost proportional to gradient of group delay waveform at current candidate calculated over a 0.55 ms window
- Autocorrelation
  - Cost proportional to negative normalized autocorrelation of speech in 10 ms window at current and previous candidates.
- Projection penalty
  - Additional cost for candidates that do not correspond to a group delay zero-crossing

The group delay slope at the zero crossing will be -1 for a pure impulse but its magnitude is less than this if there is noise in the residual waveform.

The slope therefore is a good measure of the fraction of energy in the closure impulse. If the slope is less steep then it means that the impulse has less energy and is therefore less reliable.

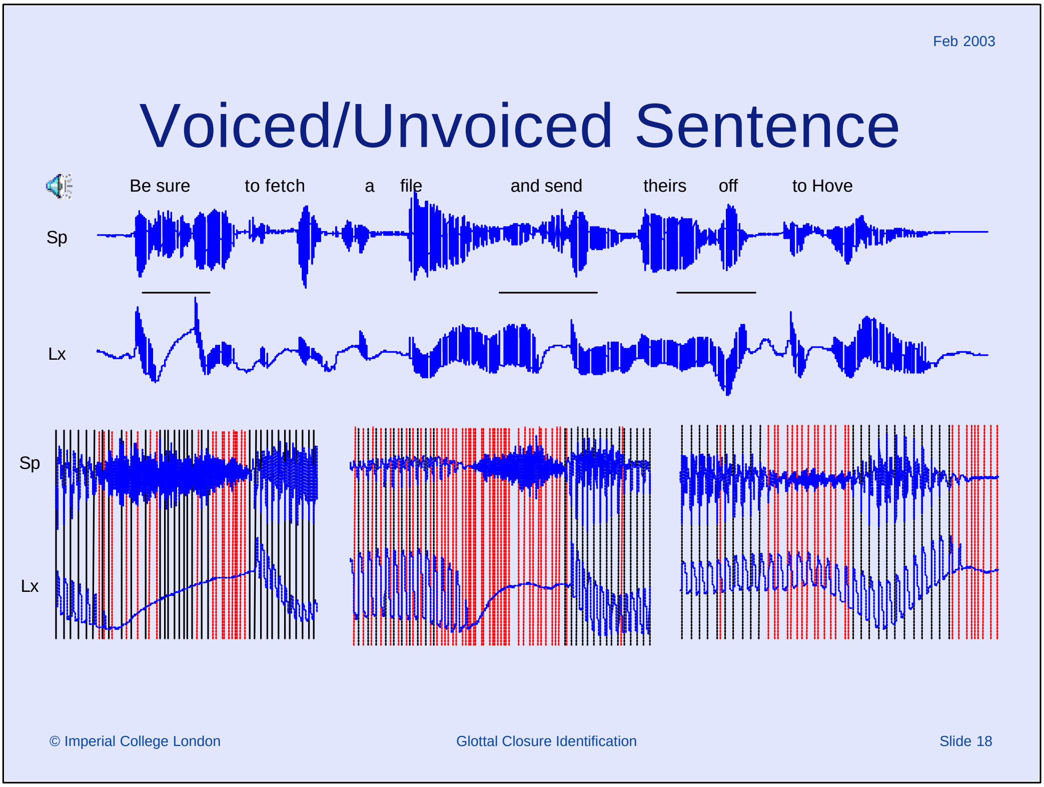
The autocorrelation term looks to see if the speech waveform is a similar shape in the previous cycle

The projection penalty gives an additional cost to candidates that did not correspond to zero-crossings.



Continuously Voiced Sentence

- Continuously voiced sentence – in the enlarged sections, the dark lines indicate the selected closure instants, the light dotted lines are the rejected candidates.
- At onset, we use an energy-based voicing detector and it has missed the first few larynx cycles. The closures have been found by stage 1 though.
- In the second segment, only one candidate per cycle.
- In the third segment, most cycles include an additional candidate at glottal opening. The DP has correctly identified the closures.
- In the final segment, at voice offset, the voicing detector has shut off too early.
- Within the central portion, there are no errors.



Voiced/Unvoiced Sentence

•With a mixed voiced/unvoiced sentence there are lots of problems at onset/offset.

•In the first segment, the /sh/ of “sure” lots of erroneous closures are found – it corrects itself as soon as voicing starts.

•In the middle segment, it has shut off too early. Lots of spurious candidates in the unvoiced region.

•In the third segment, the voicing detector has wrongly decided that it is unvoiced.

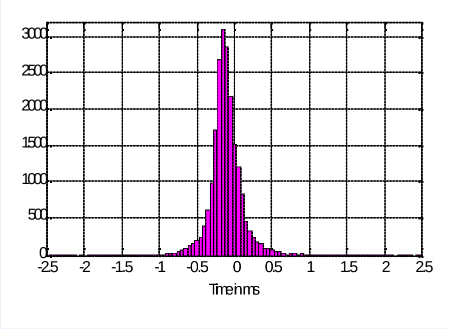
•We need to integrate the voicing decision into the DP.

Performance Assessment

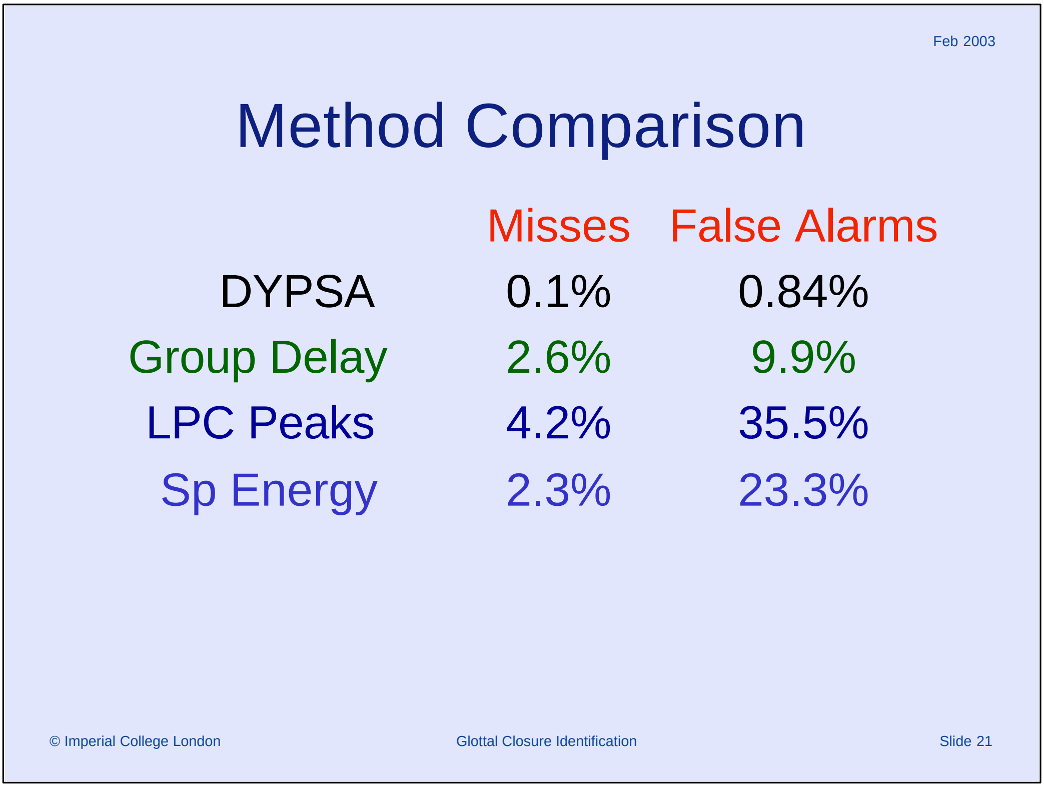
- Ignore onset/offset effects
  - concentrate on central portions of voiced segments
- Align with Lx-derived closures
  - can trade-off alignment error versus number of misses
- Ignore mean error
  - Mean timing error depends on larynx-to-microphone distance compensation

Accuracy

- Fully voiced sentence: “Why are you early you owl?”
- Start and end trimmed to avoid onset/offset disputes
- 5 utterances × 10 speakers (5m+5f)



- False alarms = 0.84%
- Missed closures = 0.1%
- Standard Deviation of matched closures = 0.29 ms



Method Comparison

	Misses	False Alarms
DYPSA	0.1%	0.84%
Group Delay	2.6%	9.9%
LPC Peaks	4.2%	35.5%
Sp Energy	2.3%	23.3%

•Group Delay method just uses zero crossings – gets 97% of closures.

•The phase slope projection is what gets us the remaining 3% although we get many extras that must be removed by DP.

•False alarm rate can always be reduced by DP – essential to start with the complete set.

•Sp Energy method uses speech energy in 3 ms window

Outstanding Issues

- Window sizes
  - 6 ms window for group delay calculation works OK for around 70 Hz to 300 Hz
- Noise robustness
  - Small test TIMIT/NTIMIT surprisingly good.
  - 90% of matching GCIs within ±1.2 ms
- Non-modal voicing
- Voiced/Unvoiced detection
  - More work needed – integrate into DP

Thank You