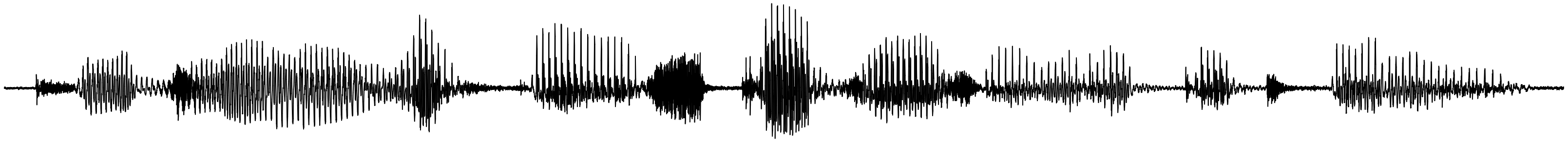
Grafeem-foneemomzetting voor spraaksynthese

Two steps

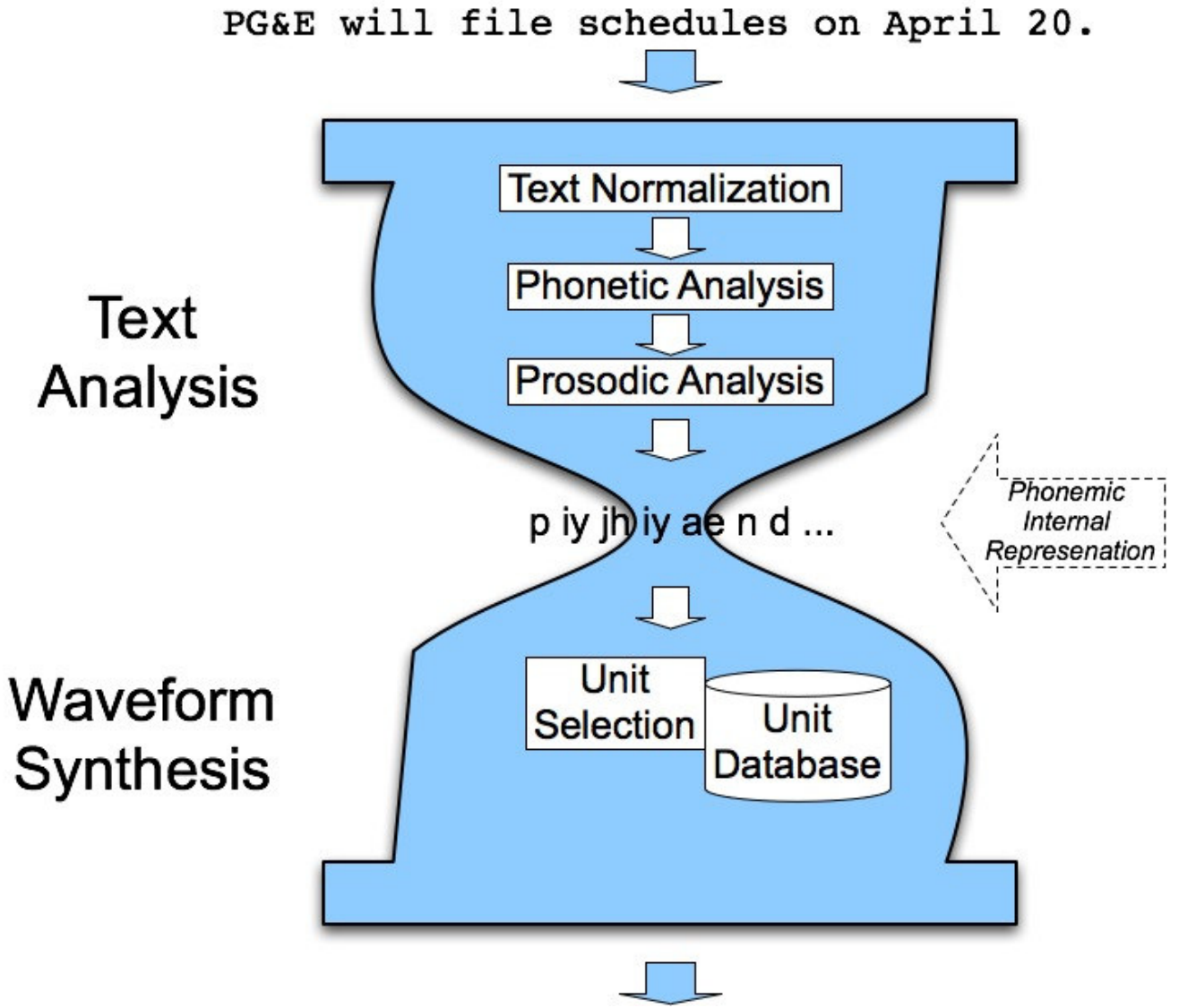
- P G & E will file schedules on April 20.
- TEXT ANALYSIS: Text into intermediate representation:

			*			*			* L-L%
P	G	AND	E	WILL	FILE	SCHEDULES	ON	APRIL	TWENTIETH
p iy	jh iy	ae n d	iy	w ih l	f ay l	s k eh jh ax l z	aa n	ey p r ih l	t w eh n t iy ax th

- WAVEFORM SYNTHESIS: From the intermediate representation into waveform



The Hourglass



fasen in tekst-naar-spraak

- Tekstbewerking
  - niet-letters, afkortingen, niet-Nederlandse woorden
- Morfologische decompositie
  - samenstellingen, syllabegrenzen [morfologie]
- Grafeem-foneemomzetting
  - lettersymbolen -> klanksymbolen [fonologie]
- Melodie en ritme (prosodie)
  - woordklemtoon [fonologie] en zinsaccenten [parsing, semantiek]
- Synthese
  - akoestische realisatie [fonetiek]

gebruikelijk ontwikkeltraject

- Dag: eerste 50%
- Week: volgende 25 %
- Maand: volgende 10 %
- Jaar: volgende 10 %
- ?: laatste 5%

tekstbewerking: cijfers

- getallen,
  5 -> "vijf", 21 -> "een en *twintig*" (twenty-one)
- telefoonnummers
  020-6680470 -> "nul *twintig* zes zes *tachtig* vier *zeventig*"
- geldbedragen
  euro 32.54 -> "twee en dertig *euro* vier en vijftig"
- huisnummers
  13-hs -> "dertien huis"
- Romeinse cijfers
  MDCCLXIV -> "zeventien honderd vier en zestig"

tekstbewerking: afkortingen

- als losse letters

  KLM -> "K L M"

- als woord

  VARA -> "vara"

- als afgekort woord

  tel. -> "telefoon"

  maar let op de juiste interpretatie van de punt
  (telefoon vs tel – *de laatste tel.*)

tekstbewerking: leestekens

- punten, komma's, puntkomma, etc -> prosodie

foneem-grafeemomzetting

symbolen voor spraakklanken

- fonetisch alfabet

  IPA International Phonetic Alphabet

  - lastige tekenset in pre-Unicode tijdperk
    œ ø ʌ ʒ ʧ ʤ ə …

- computer fonetisch alfabet

  verschillende voorbeelden, bv SAMPA

  (speech assessment methods phonetic alphabet)

spreek uit en luister naar je klanken!

3. Letter-to-Sound: Getting from words to phones

- Two methods:
  - Dictionary-based
  - Rule-based (Letter-to-sound=LTS)
- Early systems, all LTS
- MITalk was radical in having huge 10K word dictionary
- Now systems use a combination

Pronunciation Dictionaries: CMU

- CMU dictionary: 127K words
  - http://www.speech.cs.cmu.edu/cgi-bin/cmudict
- Some problems:
  - Has errors
  - Only American pronunciations
  - No syllable boundaries
  - Doesn't tell us which pronunciation to use for which homophones
    - (no POS tags)
  - Doesn't distinguish case
    - The word US has 2 pronunciations
      - [AH1 S] and [Y UW1 EH1 S]

Pronunciation Dictionaries: UNISYN

- UNISYN dictionary: 110K words (Fitt 2002)
  - http://www.cstr.ed.ac.uk/projects/unisyn/

```
going:       { g * ou }.> i ng >
antecedents:   { * a n . tˆ i . s ˜ ii . d n! t }> s >
dictionary:    { d * i k . sh @ . n ˜ e . r ii }
```

- Benefits:
  - Has syllabification, stress, some morphological boundaries
  - Pronunciations can be read off in
    - General American
    - RP British
    - Australia
    - Etc
- (Other dictionaries like CELEX not used because too small, British-only)

Dictionaries aren't sufficient

- Unknown words (= OOV = "out of vocabulary")
  - Increase with the (sqrt of) number of words in unseen text
  - Black et al (1998) OALD on 1st section of Penn Treebank:
  - Out of 39923 word tokens,
    - 1775 tokens were OOV: 4.6% (943 unique types):

names	unknown	Typos/other
1360	351	64
76.6%	*19.8%*	*3.6%*

- So commercial systems have 4-part system:
  - Big **dictionary**
  - **Names** handled by special routines
  - **Acronyms** handled by special routines (previous lecture)
  - Machine learned **g2p** algorithm for other unknown words

Names

- Big problem area is names
- Names are common
  - 20% of tokens in typical newswire text will be names
  - 1987 Donnelly list (72 million households) contains about 1.5 million names
  - Personal names: McArthur, D'Angelo, Jiminez, Rajan, Raghavan, Sondhi, Xu, Hsu, Zhang, Chang, Nguyen
  - Company/Brand names: Infinit, Kmart, Cytyc, Medamicus, Inforte, Aaon, Idexx Labs, Bebe

Names

- Methods:
  - Can do morphology (Walters -> Walter, Lucasville)
  - Can write stress-shifting rules (Jordan -> Jordanian)
  - Rhyme analogy: Plotsky by analogy with Trostsky (replace tr with pl)
  - Liberman and Church: for 250K most common names, got 212K (85%) from these modified-dictionary methods, used LTS for rest.
  - Can do automatic country detection (from letter trigrams) and then do country-specific rules
  - Can train **g2p** system specifically on names
    - Or specifically on types of names (brand names, Russian names, etc)

Letter-to-Sound Rules

- Earliest algorithms: handwritten Chomsky+Halle-style rules:

$$c \rightarrow [\text{k}] \;/\; \underline{\quad} \;\{\text{a,o}\}\text{V} \qquad ; \text{context dependent}$$

$$c \rightarrow [\text{s}] \qquad ; \text{context independent}$$

- Festival version of such LTS rules:
- (LEFTCONTEXT [ITEMS] RIGHTCONTEXT = NEWITEMS)
- Example:
  - (# [c h] C = k)
  - (# [c h] = ch)
- # denotes beginning of word
- C means all consonants
- Rules apply in order
  - “christmas” pronounced with [k]
  - But word with ch followed by non-consonant pronounced [ch]
    - E.g., “choice”

symbool omzetting

- 1 naar 1

  p -> p , a -> A

- 1 naar 2

  o -> O~

- 2 naar 1

  (dus onderzoek altijd eerst context!)

  ch -> x , sj -> S , ee -> e , ie -> i , ng -> N

- 2 naar 2

  ij -> Ei

contekst afhankelijkheid

- positie in woord

omringende letters

ban -> bAn bang -> bAN

final devoicing

dak -> dAk bad -> bAt

woordbegin <-> woordeinde

lat -> lAt tal -> tAL

wit -> wIt ruw -> ryW

assimilatie

zakdoek -> zAgduk en geen zAkduk

Voorbeelden

C*V*C* woorden zonder diacrieten

- positie van letter in woord
- positie van letter in cluster

- onset-nucleus-coda identificatie helpt
  - Klinkerclusters
    - l-eeu-w, g-eëe-rd
  - Specifieke medeklinkerclusters
    - Fonotactische restricties ->
      ook voor syllabe**gr**enzen, maar vgl we**gr**ennen

Contextgevoeligheid (C) *in C*V*C**

- Ongevoelig *in monosyllabe*
  - f, k, m, p, r, t, v, x, z
- Volgende context
  - b (#), c (varia), d (#), g (**goa**), j (#), l (#), n (**ng**, #), q (**qu**), s (**sj**), w (#)
- Vorige context
  - h (**ch**), j (**ij**), g (**ng**) vet als segment nemen

Contextgevoeligheid (V) *in C*V*C**

Veel op te lossen door segmentdefinitie

- aai, eeu, ooi, ieu
  - mooi-er , ui-en
- aa, ee, oo, uu, ie, oe, eu [1 symbool]
  ai, oi, au, ou, ij, ei, ui [2 symbolen]
  ua, io [grens uV* iV*]
- korte klinkers a, i, o, u, y
  afhandelen na lange klinkers en tweeklanken
  - kan, act = uitzondering
- e
  - lastig: e, E, @, I

Regelvolgorde

mooier

- eerst langste cluster: ooi

  m-ooi-e-r -> moj@r (mO:j@r)

- en niet

  m-oo-ie-r -> moir

  of

  m-o-o-i-e-r -> mOOIEr

	plofklanken	**wrijfklanken**	**nasalen**
labiaal (lippen)	p (**p**ad) b (**b**ad)	f (**f**iets) v (**v**at)	m (**m**at)
alveolair (achter de tanden)	t (**t**ak) d (**d**ak)	s (**s**ap) z (**z**at)	n (**n**at)
palataal (verhemelte)	tj (po**tj**e) dj (**dj**intan)	S (**sj**aal) Z (planta**g**e)	nj (a**nj**er)
velair (achter)	k (**k**at) g (za**k**doek, **g**oal)	x (la**ch**) G (**g**at**)**	N (ba**ng**)
glottaal (stem)	_ (stilte)	h (**h**uis)	

Liquida en halfvokalen

	woordbegin	**woordeinde**
Liquida	l (**l**ang, a**ll**e) r (**r**ug)	L (a**l**) r (haa**r**) R (haa**r**)
Halfvokalen	w (**w**it) j (**j**an)	W (sneeu**w**) na / i, e, y / J (aa**i**)

Korte klinkers

	voor		**achter**	
	[-rond]	**[+rond]**	**[-rond]**	**[+rond]**
hoog	I (b**i**d)	Y (p**u**t)		
midden	E (b**e**d)		@ (red**e**)	O (b**o**s)
laag			A (b**a**k)	

Lange klinkers

	voor		**achter**	
	[-rond]	**[+rond]**	**[-rond]**	**[+rond]**
hoog	i (b**ie**d)	y (b**uu**t)		u (b**oe**k)
midden	e (b**ee**t)|	2 (b**eu**k)		o (b**oo**t)
laag			a (b**aa**t)	

Tweeklanken

	voor		**achter**	
	[-rond]	**[+rond]**	**[-rond]**	**[+rond]**
midden	Ei (b**ij**t)	9y (b**ui**t)		Au (b**ou**t)

Gekleurde klinkers voor r

	voor		**achter**	
	[-rond]	**[+rond]**	**[-rond]**	**[+rond]**
midden	I: (b**ee**r)	Y: (d**eu**r)		O: (d**oo**r)

Franse klanken

	voor		**achter**	
	[-rond]	**[+rond]**	**[-rond]**	**[+rond]**
	E~ (manneq**ui**n)	Y~ (parf**u**m)	A~ (ch**a**nson)	O~ (chans**o**n)

[Oe] uit fr**eu**le