Building Voices in The Festival Speech Synthesis System:

The *festvox* project

Alan W Black
Language Technologies Institute
Carnegie Mellon University
(this is joint work with Kevin Lenzo, CMU)

Overview

- ☐ Goal:
  - automatically build new synthesis voices
- ☐ **festvox** provides:
  - documentation
  - scripts, tools
  - autolabeling, model building and diagnosis
  - `http://www.festvox.org/`

Results may not be *best* quality
– but can still be *good* quality.

Festival as software

- General system for multi-lingual TTS
- C/C++ code with Scheme scripting language
- General replaceable modules:
  - lexicons, LTS, duration, intonation, phrasing, POS tagging, tokenizing, diphone/unit selection, signal processing
- General tools:
  - intonation analysis (F0, Tilt), signal processing, CART building, n-gram, SCFG, WFST, OLS
- No fixed theories
- New languages without new C++ code
- Multiplatform (Unix/NT)
- Full sources in distribution
- Free software

Why provide this?

- Forces expertise to be documented
- Forces models to be *genuinely* generic
- Increases knowledge base
- Allows speech technology access for all
- There is a demand
  - *"I want it to talk like me."*
  - or (release it and there will be a demand)

The “standard” voice requires ...

Relative difficulty of parts is language dependent

- Text analysis:
  - number, symbol expansion, etc
- Pronunciations:
  - lexicon/letter-to-sound rules
- Prosodic modelling:
  - phrasing, intonation, duration etc
- Waveform generation:
  - diphones or unit selection
- *plus* something else that's difficult:
  - e.g. word segmentation in Chinese
  - vowels in Arabic
  - number declensions in Slavic languages

Text analysis

- This is a pen.
- My cat who lives dangerously has nine lives.
- He stole $100 from the bank.
- He stole 1996 cattle on 25 Nov 1996.
- He stole $100 million from the bank.
- It's 13 St. Andrew St. near the bank.
- Its a PII 450Mhz, 128MB RAM, 4.5Gb SCSI, 24x cdrom and 19" monitor.
- My home pgae is `http://www.geocities.com/awb/`.

Text analysis

traditionally *"bunch of hacky rules"*

But here, based on NSW Project from JHU Workshop '99
(Analysis of non-standard words)

- splitter:
  - domain independent tokenizer
- token type classifier:
  - number (ordinal/cardinal/year) etc
  - homograph disambiguator
  - abbrev/letter sequence identifier
- token type expander:
  - mostly deterministic
  - but abbreviation expander interesting
- language model:
  - (optionally) choose best expansion

NSW model for new domains

- Models for specific domains
- Standard text analyzers fail
- Can build models from unlabelled data

57 ST E/1st & 2nd Ave Huge
drmn 1 BR 750+ sf, lots of sun &
clsts. Sundeck & lndry facils. Askg
$187K, maint $868, utils
incld. Call Bkr Peter 914-428-9054.

Lexicons and letter to sound rules

- Need big list of words:
  – often hardest requirement
- Sharing between dialects:
  – CSTR English dialect lexicon
- But the lexicon is *never* complete:
  – need out of vocabulary pronouncer

Dialect synthesis

- ☐ Different dialects have different styles:
  – phones, prosody, vocab and grammar
- ☐ Build lexicon with "key vowels":
  – the vowel in *coffee*
- ☐ Record person's "coffee" vowel ...

Dialect synthesis:
– RAB, British RP English
– JGM, Southern Irish English
– KED, Standard US English

– I say Pakistan (RAB,JGM)
– What a waste (RAB,JGM)
– Its hotter in the city (JGM,KED)

Letter to sound rules

- Writing rules by hand is difficult
- We provide automatic process:
  - built from lexicon
  - provides phone string plus stress
  - tested on multiple languages

Lexicon	Correct	
	Letters	Words
OALD	95.80%	74.56%
CMUDICT	91.99%	57.80%
BRULEX	99.00%	93.03%
DE-CELEX	98.79%	89.38%

Letter to sound rules: method

- Find alignments:

c	h	e	c	k	e	d
ch	_	eh	_	k	_	t

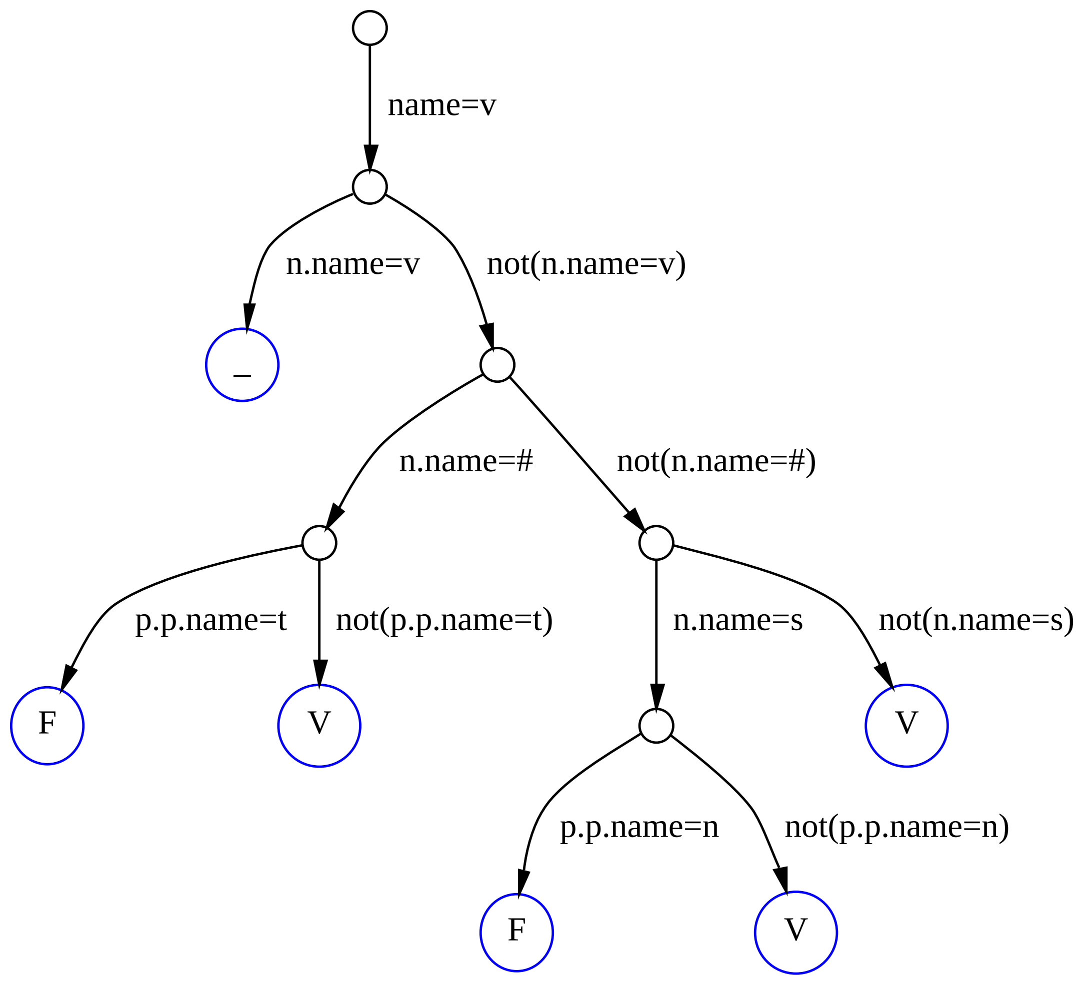
  – Using epsilon scattering and EM algorithm
  – or hand specify possible letter-phone combinations

- Build CART models:
  – predict phone (or epsilon)
  – letter context (plus POS ...)

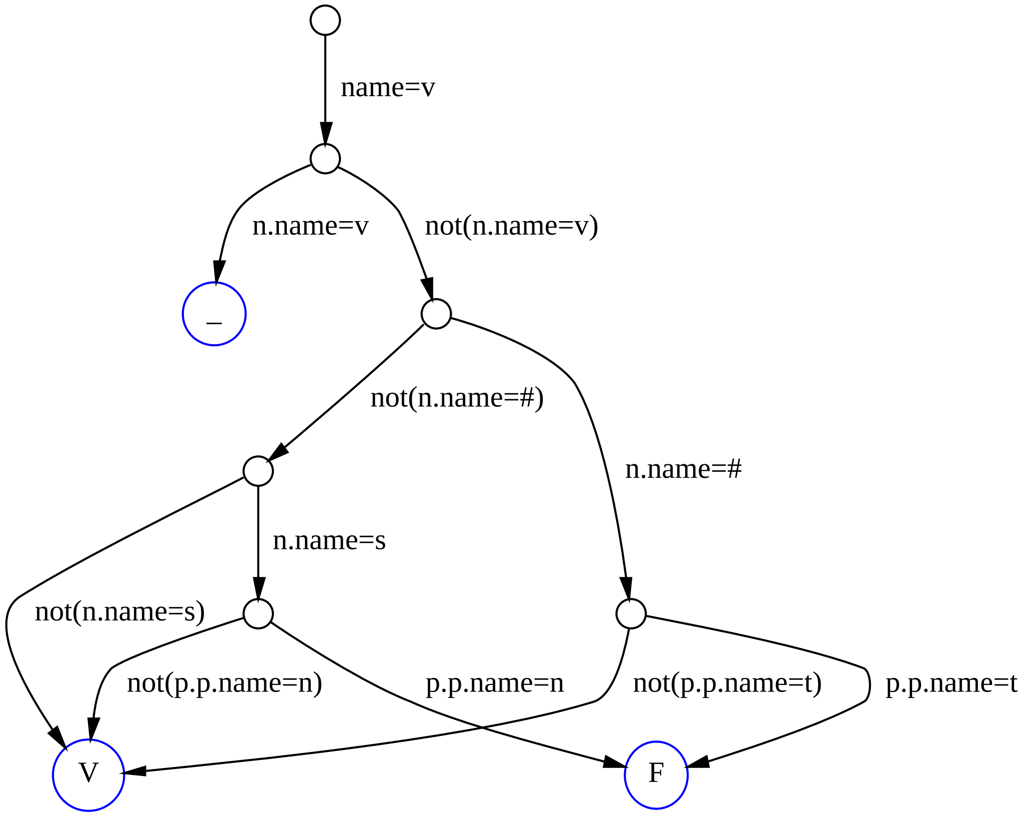
An example tree

```
For letter V:
if (n.name is v)
  return _
  if (n.name is #)
      if (p.p.name is t)
        return f
        return v
      if (n.name is s)
        if (p.p.p.name is n)
          return f
          return v
        return v
```

Vtree as a graph



Vtree as a minimized graph



Minimizing CART models

- Treat trees as FSMs
- Arcs labelled with questions
- Standard determinize/minimized algorithms
- Reduced trees to minimized graphs

	Trees	Mingraphs
OALD	14968	6339

Could optimize graph with multi-questions

Prosody modelling

phrasing, accent prediction, duration and F0

- ☐ Simple hand written rules:
  – and not so simple hand written rules
- ☐ Cross language modelling:
  – e.g. English duration model "works" for German
- ☐ Data driven techniques:
  – database building aids
  – extraction and statistical modelling
  – e..g CART accents, LR F0 modelling
  – But requires well labelled dbs
- ☐ Model adaptation (for supported languages):
  – take small example and adapt larger model
  – cf. Shih, Gu and van Santen, ESCA 98.

Waveform synthesis: diphone

- Design schema:
  – nonsense words
- Synthesize them as prompts:
  – can (sort of) be done cross-language
  – as monotone, one or two diphones per word
- Prompt and record:
  – prompts help normalize speech
  – prompts save mistakes
- DTW align prompt with spoken words:
  – Malfrere and Dutoit Eurospeech 97
- Extract pitchmarks:
  – from EGG signal or direct from waveform
- LPC extraction
- *Hand* correction

... or just use an existing MBROLA database.

diphone example: (US English)

- Diphone schema 1348 words (1618 diphones):
  – includes consonant clusters and open vowels
- Prompts are:

```
kal_0001 ("b-aa" "aa-b") (t aa b aa b aa)
kal_0002 ("p-aa" "aa-p") (t aa p aa p aa)
```

- Automatic aligning alone:
  – around 15 words *wrong*
  – better than human labeller (first pass)
  – takes 30 minutes (vs 2 weeks)
- After tidy up (couple of hours)
  – example: fully automatic
  – example: with correction
- KAL voice created in 2 days

Voice building

- Completed:
  – English (US, UK (RP, Scots and Irish), Oz)
  – German (4), Spanish (Castillean and Mexican), Welsh, Polish, Greek, Portuguese, Swedish, Korean, Basque
- In progress (that I know of):
  – Turkish, Maori, Chinese, Hindi, Indian English, Japanese, Serb-Croat

Unit selection

Cluster unit selection (Black and Taylor Eurospeech 97)

- ☐ Decision tree partitioning on acoustic distance:
  – phone index (but diphone selection)
  – pitch synchronous melcep plus power and F0
  – indexed by phone and prosodic context
  – cf. Donovan 95
- ☐ Viterbi selection through candidates
  – cf. Hunt and Black 96, Iwahashi and Sagisaka 95
- ☐ Does it work?
  – good in parts

Taylor and Black 99, Phonological Structure Matching not yet integrate into distribution.

Limited Domain Synthesis

For (currently) *closed* but infinite domains:
– e.g. times, telephone numbers, weather (?)

- ☐ Data files of utterances:

```
( time0001 "The time is now, exactly five past one, in the morning." )
( time0002 "The time is now, just after ten past two, in the morning." )
...
( time0023 "The time is now, exactly five past eleven, in the evening." )
( time0024 "The time is now, a little after quarter to midnight." )
```

- ☐ Prompt and record
- ☐ Automatically align
- ☐ Build clunits synthesizer (overkill)
- ☐ Synthesize any time ...

Ending comments

- Basic aids do work
- Non-Edinburgh/OGI/CMU groups have success
- Second release of **festvox** imminent
- Much work still to be done:
  - autolabeling (error detection/correction)
  - easier prosodic modelling
  - NSW text analysis packaging
  - Testing and evaluation support
- Creating a larger community