Seminario 3. Síntesis de voz con Festival

Sistemas Multimedia Interactivos e Inmersivos

Grado de Ingeniero en Informática

Escola Tècnica Superior de Enginyeria Informàtica

Curso 2018/2019

Manuel Agustí

Índice

Introducción a Festival

- Tema 2. Audio
  - *Text to Speech* (TTS)
  - Multiplataforma
  - Estándar
  - Abierto

- Aplicaciones / SDK
  - Festival /Flite /Flinger
  - Espeak <http://espeak.sourceforge.net/>
  - ...?

Introducción a Festival (II)

- *The Festival Speech Synthesis System*
  - *Fextvox*
    - *Carnegie Mellon University's speech group*
  - *Edinburgh Speech Tools*
    - *O*peraciones básicas: EST_*
    - The Centre for Speech Technology Resea
      The university of Edinburgh
  - Intérprete de órdenes (SIOD)
  - API para C++ y Java
  - *Ejemplos:*
    - *Online demo*
    - *Technical online demo with more voices*

- *Festival Text-to-Speech Online Demo*

Festival Text-to-Speech Online Demo

Select a Voice

Alan (Edinburgh male)

Type the text to synthesise (max 70 chars)

Type your text here.

say it!

This is an interactive demo of CSTR's "Festival" speech synthesiser, which is software capable of making artificial speech in place of a real human. Festival is the most complete freeware multilingual, general-purpose synthesis system available. It is used by numerous research sites and other projects around the world. Further information is available on the Festival project page.

Synthetic speech can be used anywhere pre-recorded speech can, for example telephone call routing and information lines. However, it is equally useful where pre-recorded speech **cannot** be used. For example, screen-readers for the visually impaired, email/sms reading over the telephone, voice directions for in-car satellite navigation and gaming dialog to name but a few.

The following voices are included in this demo at present:

- Alan - Scottish male.
- Nick - English RP male.
- Roger - English RP male
- Nina - English RP female.
- KAL - American male.
- SLT - American female.

- *Festival Text-to-Speech Online Demo - Technical*

The Centre for Speech Technology Research
The University of Edinburgh

Home | People | News | Research | Publications | Opportunities | Downloads | Contact

[festival] [intranet]

Festival Text-to-Speech Online Demo - Technical

Select a Voice

Nick - 1 (English RP)

Type the text to synthesise (max 70 chars)

say it!

This is an interactive demo of CSTR's "Festival" speech synthesiser, which is software capable of making artificial speech in place of a real human. Festival is the most complete freeware multilingual, general-purpose synthesis system available. It is used by numerous research sites and other projects around the world. Further information is available on the Festival project page.

Unlike the simpler demo here, the demo on this page gives access to many more voices which have been developed for Festival. This is intended to allow closer scrutiny of the results of different synthesis methods and different subsystems at various stages of development. The following voices are included at present, with an indication of the amount of speech data used to build the voice:

- Scottish male - Alan (ARCTIC), Jon (2hr)
- English RP male - Nick (8hr), Roger (13hr), Korin (TIMIT, ~20mins)
- English RP female - Nina (3hr)
- American male - KAL (Communicator), RMS (ARCTIC), BDL (ARCTIC), JMK (ARCTIC)
- American female - SLT, CLB (both ARCTIC)

Broadly, three synthesis methods are available in this demo:

- HTS - a statistical parametric approach (both the 2005 and 2007 systems)
- Multisyn - standard unit selection concatenative approach
- Diphone - single instance diphone concatenation
  (the previous TTS generation technology, from mid 1980's to mid 1990's).

Introducción a Festival (III)

- *¿Diferencias entre las dos "demos"?*
  - *Métodos de síntesis disponibles*
    - *HTS - a statistical parametric approach (both the 2005 and 2007 systems)*
    - *Multisyn - standard unit selection concatenative approach*
    - *Diphone - single instance diphone concatenation (the previous TTS generation technology, from mid 1980's to mid 1990's).*
      - *festvox_kallpc16k.tar.gz American English male voice `kal' 16kHz*
      - *festvox_kallpc8k.tar.gz American English male voice `kal' 8kHz*
      - *festvox_kedlpc16k.tar.gzAmerican English male voice `ked' 16kHz*
      - *festvox_kedlpc8k.tar.gz American English male voice `ked' 8kHz*
      - *festvox_rablpc16k.tar.gzBritish English male voice `rab' 16kHz*
      - *festvox_rablpc8k.tar.gz British English male voice `rab' 8kHz*

Arquitectura de un sistema TTS

- *TTS = front-end + back-end*

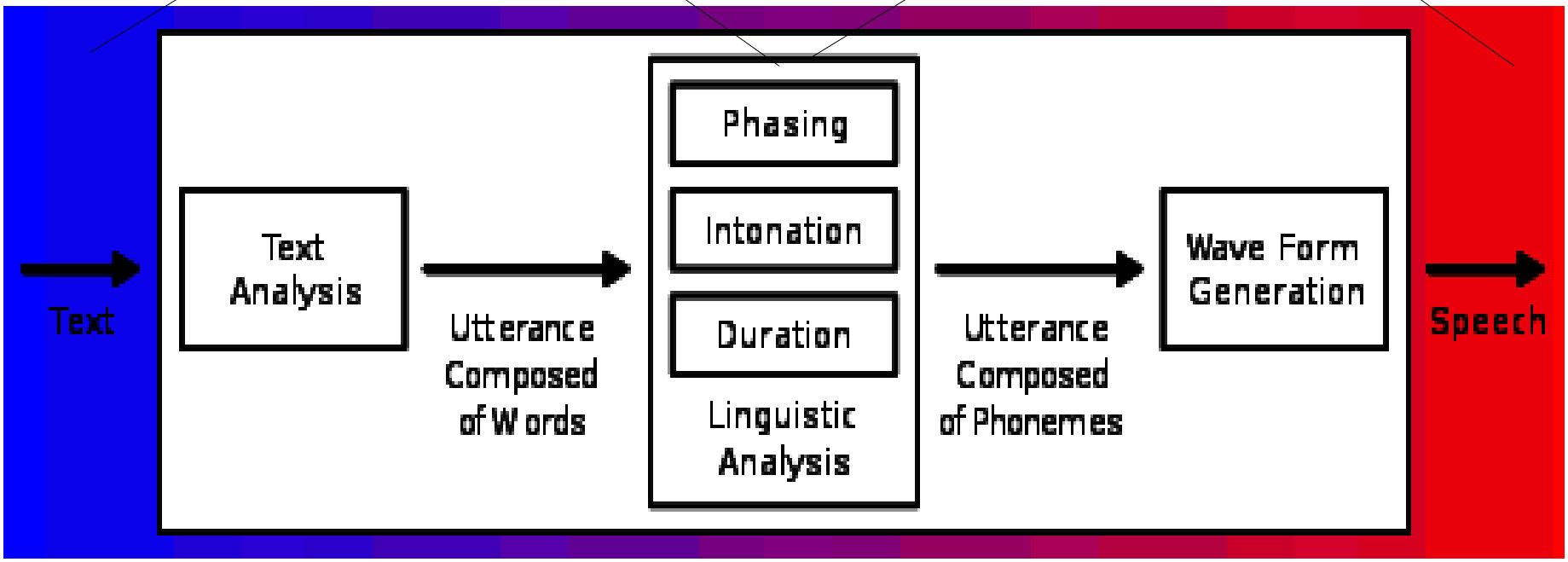
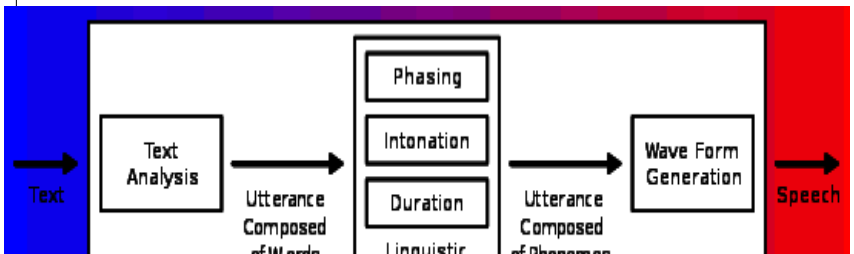


Imagen de <http://en.wikipedia.org/wiki/Speech_synthesis>

- *Partes:*
  - Análisis del texto (preproceso / normalización)
  - Análisis lingüístico (descripción fonética)
  - Generación del sonido

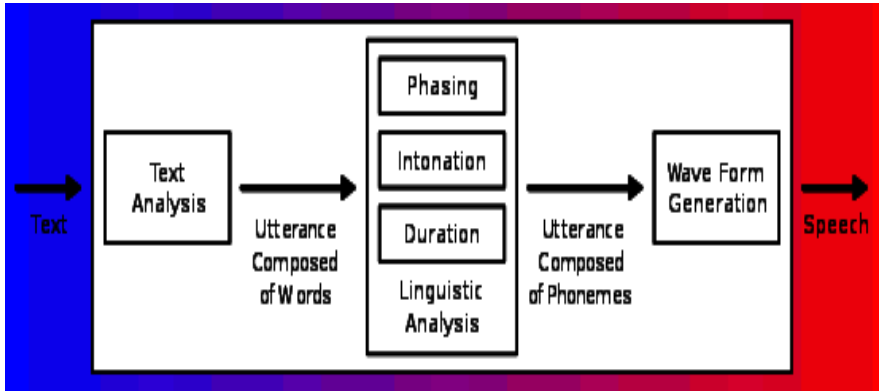
Arquitectura de un sistema TTS

- Análisis del texto (preproceso / normalización)
  - texto → a la identificación de palabras, signos de ortografía y elementos básicos del habla (números y cifras, abreviaturas, siglas, ….)
  - *Utterance chunking* is an externally specifiable part of Festival ← it may vary from language to language
    - Tokens = espacios ==> *utterances* (puntos, interrogantes, exclamaciones, ...)
    - Japonés/Chino no suelen utilizar espacios en blanco puntuación → “utterance boundaries”



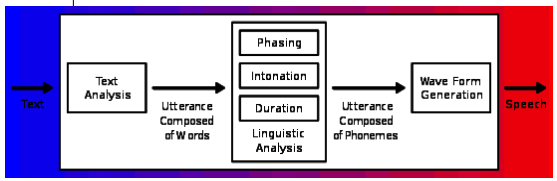
Arquitectura de un sistema TTS

- Análisis lingüístico (descripción fonética)
  - A partir de la determinación de las palabras y los signos de puntuación → Entonación/prosodia
    - Se observa como cambios en *pitch*, volumen, cualidades de la voz o velocidad de "reproducción" del mensaje
    - En contextos abiertos es muy difícil recoger suficientes muestras para aprenderla
      - Conjunto de características que se modelan estadísticamente
        - phrasing, duration, intonation
        - energy , voice quality



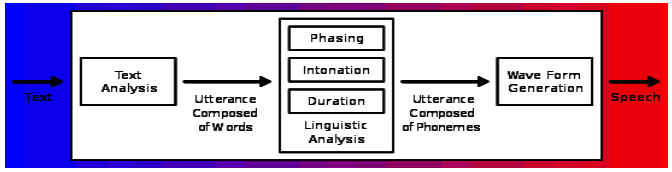
Arquitectura de un sistema TTS

- Generación del sonido
  - *From a fully specified form (pronunciation and prosody) generate a waveform.*
  - Métodos
    - Síntesis de sonidos (FM y síntesis aditiva)
    - Concatenativos
      - Grabaciones de voces reales (*voice quality*).
      - Diccionarios de n-fonemas (fonemas, sílabas, palabras, frases y oraciones).
    - Articulatorios
      - Modelo del sistema articulatorio de la voz humana
      - Diccionario de pronunciación
    - HMM
      - Modelo estadístico de síntesis paramétrica



Arquitectura de Festival

- *Arquitectura de Festival*
  - *TTS*
  - Además
    - Eficiencia del sistema de audio
      - *spooling audio files while the rest of the synthesis process can continue.*
    - Permite incorporar nuevas voces al exportar los mecanismos de
      - *basic utterance structure, a language to manipulate it, and methods for construction and deletion;*
      - *basic analysis tools (pitch trackers, classification and regression tree builders, waveform I/O etc) : Edinburgh Speech Tools (EST)*
      - *and a simple but powerful scripting language (SIOD)*
    - *FestVox, MBROLA, ...*



Introducción a la práctica de Festival

- Modos de funcionamiento
  - Intérprete de órdenes *Scheme* (SIOD)
  - Línea de órdenes
  - Código C++

Introducción a la práctica de Festival

- Modos de funcionamiento
  - Intérprete de órdenes *Scheme* (SIOD)

```
$ festival
Festival Speech Synthesis System 2.1:release November 2010
Copyright (C) University of Edinburgh, 1996-2010. All rights reserved.

clunits: Copyright (C) University of Edinburgh and CMU 1997-2010
hts_engine:
The HMM-based speech synthesis system (HTS)
hts_engine API version 1.04 (http://hts-engine.sourceforge.net/)
Copyright (C) 2001-2010  Nagoya Institute of Technology
            2001-2008  Tokyo Institute of Technology
All rights reserved.
For details type `(festival_warranty)'
festival>
festival>  (voice.list)
…
festival> (SayText "Hola, mundo!")
festival> language_default
english
festival> language-path
("/usr/share/festival/languages/")
festival> voice_default
voice_kal_diphone
festival> voice-path
("/usr/share/festival/voices/")
festival> (voice_JuntaDeAndalucia_es_sf_diphone)
…
festival> (quit)
```

```
(rab_diphone
 don_diphone
 ked_diphone
 kal_diphone
 JuntaDeAndalucia_es_pa_diphone
 el_diphone
 JuntaDeAndalucia_es_sf_diphone
 lp_diphone
 pc_diphone)
```

Introducción a la práctica de Festival

- Modos de funcionamiento
  - Línea de órdenes

```
$ echo “Hola, mundo” | festival –tts
$ echo "(voice.list)" | festival -i

$ echo “Hola, mundo” | festival --tts --language spanish

$ echo "muchísimas castañas" | festival --tts --language spanish
$ echo "muchísimas castañas" | iconv -f utf-8 -t iso-8859-1 | festival --tts --language spanish


$ text2wave texto.txt -otype snd -o habla.wav

$ festival fichero.sable --tts
```

Introducción a la práctica de Festival

- Festival
  - Código C++

```
festival_initialize(load_init_files,heap_size);
festival_say_file(const EST_String &filename)
festival_say_text(const EST_String &text)
festival_eval_command(const EST_String &expr)
festival_load_file(const EST_String &filename)
festival_text_to_wave(const EST_String &text,
                      EST_Wave &wave)
festival_wait_for_spooler()
```

```
#include <stdio.h>
#include <festival.h>

int main(int argc, char **argv)
{
  int heap_size = 210000;
  int load_init_files = 1;

 festival_initialize(load_init_files, heap_size);

 festival_eval_command( "(voice_el_diphone)" );
 festival_say_text( "Hola mundo" );
}
```

```
$ g++   ejemplo.c -o ejemplo -lFestival -leststring   -lestbase   -lestools
```

Otras posibilidades de Festival

- Sable

```
<?xml version="1.0"?>
<!DOCTYPE SABLE PUBLIC "-//SABLE//DTD SABLE speech mark up//EN"
      "Sable.v0_2.dtd"
[]>
<SABLE>

<!-- A basic set of examples from SABLE -->
Without style, <BREAK LEVEL="Large"/> Grace and I are in trouble.

<DIV TYPE="paragraph">
 <DIV TYPE="sentence" >
   Yesterday, Denmark and India announced an agreement of cultural exchange. </DIV> <DIV TYPE="sentence"> Further talks will take place next month.
 </DIV>
</DIV>

The leaders of <EMPH>Denmark</EMPH> and <EMPH>India</EMPH> meet on Friday.
An example is <ENGINE ID="festival" DATA="our own festival speech synthesizer"> the festival speech synthesizer</ENGINE> or the Bell Labs speech
synthesizer.

Some English first and then some Spanish.
<LANGUAGE ID="SPANISH">Hola amigos.</LANGUAGE>
<LANGUAGE ID="NEPALI">Namaste</LANGUAGE>

Move the <MARKER MARK="mouse" /> mouse to the top.
Without his penguin, <PITCH BASE="-20%"> which he left at home, </PITCH> he  could not enter the restaurant.
I say <PRON SUB="toe maa toe">tomato</PRON> and you say <PRON SUB="toe may toe">tomato</PRON>.
The address is <RATE SPEED="-40%"> 10 Main Street </RATE>.


As a test of marked-up numbers. Here we have a year <SAYAS MODE="date">1998</SAYAS>, an ordinal <SAYAS MODE="ordinal">1998</SAYAS>, a
cardinal <SAYAS MODE="cardinal">1998</SAYAS>, a literal <SAYAS MODE="literal">1998</SAYAS>, and phone number <SAYAS
MODE="phone">1998</SAYAS>
Please speak more <VOLUME LEVEL="loud">loudly</VOLUME>, except when I ask you to speak <VOLUME LEVEL="quiet">in a quiet voice</VOLUME>.

</SABLE>
```

Casos de estudio con Festival

- V. Pérez
  - E-Narrador
  - Web-a-voz
- Podría ser
  - Un reloj que habla
  - Adivina el número
  - Pronunciación de idiomas
  - ...

Bibliografía

- *V. Pérez. (2011). Estudio de los motores de síntesis del habla. PFC.*
- *The Festival Speech Synthesis System* <http://www.cstr.ed.ac.uk/projects/festival/>
- *Festvox <http://www.festvox.org>*
- MBROLA <http://tcts.fpms.ac.be/synthesis/mbrola.html>
- *Speech Synthesis & Analysis Software* <http://linux-sound.org/speech.html>
- Speech Synthesis <http://en.wikipedia.org/wiki/Speech_synthesis>
- Alicebot <http://www.alicebot.org/>

Click here to chat with Talking Animated Fake Captain Kirk

  - Eliza ('64) <http://en.wikipedia.org/wiki/ELIZA>
  - Dr. Sbaitso ('92) <http://en.wikipedia.org/wiki/Dr._Sbaitso>