Evaluación Albayzín 2008

V Jornadas en Tecnología del Habla

Red Temática en Tecnologías del Habla

Índice

- ✓Verificación de idioma
- ✓Síntesis de voz
- ✓Traducción automática

Características de la evaluación

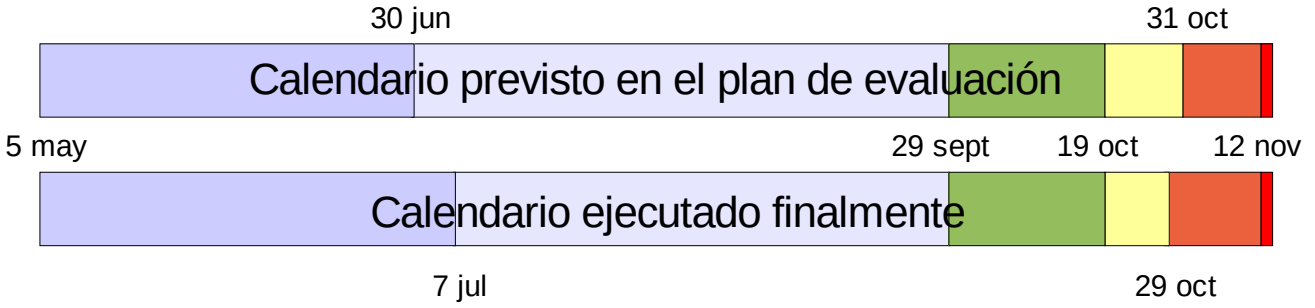
- Organización: **Luis Javier Rodríguez**
- 4 lenguas consideradas: **Castellano, Catalán, Euskera y Gallego**
- Objetivo: Dado un segmento de señal *S* y una cierta lengua *Li*, la tarea consiste en decidir si el habla contenida en *S* corresponde a la lengua *Li*
- Sistemas libres y restringidos
- En modo cerrado y en modo abierto
- **Entradas:** voz, lengua objetivo y lenguas contraste
- **Salidas:** decisión SÍ/NO y probabilidad

Descripción del Corpus KALAKA

- Grabaciones de TV y Radio: condiciones ambientales y de canal muy diversas: entrevistas en estudio sin ruido de fondo, reportajes desde la calle, ...
- Entrenamiento: 35.49 horas, Desarrollo: 7.69 horas, Evaluación: 7.68 horas (ficheros de 3, 10 y 30 segundos). Formato 16KHz (16 bits)
- Evaluación según protocolo y medidas del NIST LRE 2007: Cavg y curvas DET
- Premio: Competición CR – 30 segundos

Calendario

- Publicación del plan de evaluación e inscripción de equipos
- Desarrollo de sistemas a partir de los materiales de entrenamiento y desarrollo
- Procesamiento de los materiales de evaluación y envío de resultados de verificación
- Procesamiento y análisis de resultados por parte de la organización
- Notificación de resultados, liberación del fichero de claves y preparación de presentacione
- Workshop Evaluación ALBAYZIN08-VL: V Jornadas en Tecnología del Habla (Bilbao)



Participantes

- 4 equipos 13 sistemas
- Todas universidades

Abierto y Libre

Competición / *Equipo*	AL	AR	CL	CR
ATVS-UAM	pri	pri, alt	pri	pri, alt
L2F-INESC	pri	pri	pri	pri
PRHLT-UPV				pri, alt
SOFTLAB-UC3M				pri
Total Sistemas	*2*	*3*	*2*	*6*

Participante: ATVS-UAM

- Entrenamiento Restringido (R)
  - Sistema Primario: GMM-SVM + AGMM-SVM
    - GMM-SVM: Fusión de 2 sistemas GMM-SVM sobre distintas características: 9 MFCC+Δ y 49 SDCs (7-1-3-7), CMN + RASTA + Feature Warping
    - AGMM-SVM: GMM-SVM alternativo, con fusión a nivel de características:7 MFCCs + 49 SDCs (7-1-3-7), CMVN + RASTA
  - Sistema Alternativo: AGMM-SVM
- Entrenamiento Libre (L)
  - Sistema Primario: GMM-SVM + AGMM-SVM + Phone-SVM
    - Phone-SVM: 7 rec. fon. en 7 idiomas -> 1-gram+2-gram+3-gram -> SVM
- Calibración individual por sistema, duración, idioma y condición (A/C)
  - LogLRs obtenidos mediante Linear Logistic Regression (FoCal Toolkit)
- Fusión promedio de sistemas calibrados
- Decisión: Umbral de Bayes

Participante: L2F-INESC

GMM

- 7 static PLP with RASTA + 49 Shifted Delta (7-1-3-7)
- GMMs of 1024 mixtures (400 minutes per GMM/language)
- VQ initialization + 10 EM iterations
- Back-end normalization and scoring: T-norm with mean of three competing log-likelihoods and Z-norm language dependent estimated on 100 minutes per language.

FOUR PARALLEL PHONOTACTIC SYSTEMS: PT, BR, SP and EN

Linear SVM combination

- 4 binary SVMs (100 minutes each)
- 16-element input vector (4 PPRLM x 4 target languages)
- Different decision threshold for open/closed

Phonetic tokenizers

- Multiple-stream MLP classifiers
- PT outputs phonemes
- BR, SP and EN outputs sub-phonetic units

Phonotactics modelling and normalization

- 3-gram modes for each target language
- Use of 400 minutes of data per language
- T-norm with mean of three competing scores

Participante: PRHLT-UPV

- Sistema primario
  - Preproceso:
    - Parametrizador de audio de iATROS
    - Shifted Delta Cepstrum (SDC) a partir de cuatro parámetros, N-d-P-k
  - Sistema:
    - Modelo GMM con 4096 gaussianas
    - Entrenado con HTK y los SDCs obtenidos en el preproceso
    - Los SDCs se obtuvieron con valores de N-d-P-k de 7-1-3-7
    - Vectores de características de 49 dimensiones calculadas cada 210ms
- Sistema alternativo
  - Preproceso:
    - Cepstrales con primera y segunda derivada
  - Sistema alternativo:
    - Se aprende un codebook de C codewords con el algoritmo c-medias
    - Cuantificación vectorial de los ficheros según dicho codebook aprendido
    - Histograma de la frecuencia de aparición de cada uno de los codewords
    - Se aprende una base de proyección de C a d dimensiones
    - Dicha proyección se aplica tanto a vectores de entrenamiento como de test

Participante: SOFTLAB-UC3M

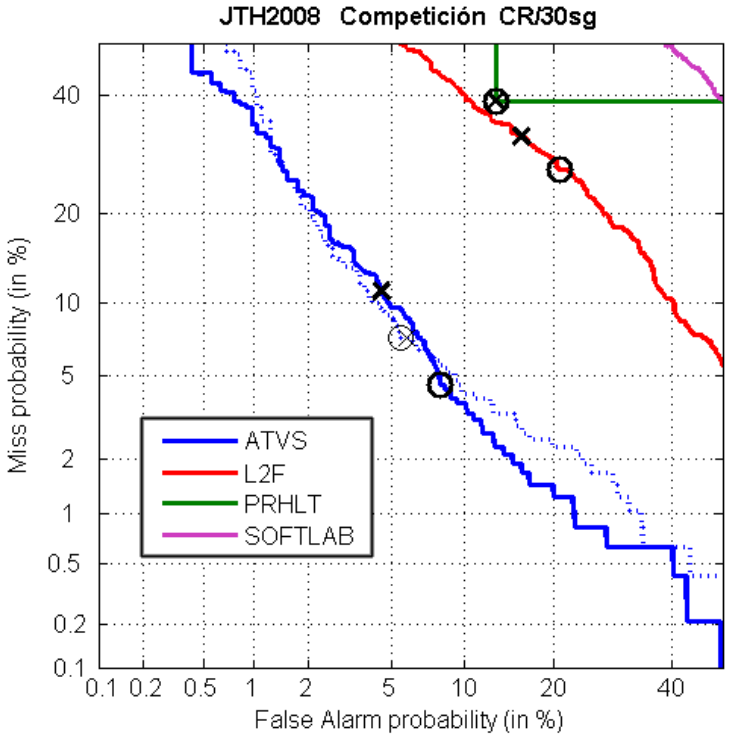
- Características
  - MFCC + E y deltas con CMN
  - Máquina de estados
    - HMMs de 6 estados
    - 4 umbrales para detectar transiciones
  - Clasificador SVM
    - Caracterización acústica de cada idioma
    - Obtención del hiperplano de separación entre los vectores pertenecientes y no al idioma.
    - Kernel RBF

Resultados

- Mejor sistema primario en la competición CR-30 : ATVS-UAM (C_{avg} = 0.0778)

	C_{avg}					
Competición	**AL-30**	**AR-30**		**CL-30**	**CR-30**	
Sistema	*primario*	*primario*	*alternativo*	*primario*	*primario*	*alternativo*
ATVS-UAM	0.0946	0.1313	0.1110	0.0552	**0.0778**	0.0656
L2F-INESC	0.1204	0.2787	--	0.0556	0.2420	--
PRHLT-UPV	--	--	--	--	0.2597	0.5389
SOFTLAB-UC3M	--	--	--	--	0.5035	--

Resultados



Características de la evaluación

- Organización: **Iñaki Sainz y Eva Navas**
- Objetivo: Generación de un sintetizador con la mayor calidad posible
- **Entradas:** frases a sintetizar
- **Salidas:** voz para esas frases
- 8 participantes: 6 universidades y 2 empresas

Descripción del Corpus

- Referencia: *Ignacio Hernández, Asunción Moreno. Diseño de un corpus para una base de síntesis de voz. Actas del XV Symposium Nacional de la U.R.S.I., Zaragoza (España), Septiembre 2000.*
- Conjuntos de entrenamiento: más de 1 hora de voz de una locutora profesional (Marta)
- Evaluación: más de 300 frases de las que se elegirán aleatoriamente las que forman parte del test.
- Medidas de evaluación: evaluación perceptual por web con un cuestionario sobre diferentes aspectos de la voz: naturalidad e inteligibilidad

Participante: Barcelona Media Center & Cereproc

- Unidad base: Difonema
- Tipo: concatenación
- Modelo prosódico: sencillo
- C. Objetivo: Acento, límites pros.,F0 y duración.
- C. Concatenación: LSF, F0 y energía
- R. espectral: LSF.
- Otras: ESPS para obtener el pitch

Participante: UPM

- Unidad base: Pentafonema
- Tipo: HMMs
- Modelo prosódico: HMMs
- C. Objetivo: -
- C. Concatenación: -
- R. espectral: Mel Cepstrum
- Otras: Modelo de excitación mixta

Participante: Telefónica I+D

- Unidad base: Difonema
- Tipo: Concatenación
- Modelo prosódico: estadístico y selección por corpus
- C. Objetivo: sonidos, contexto y prosodia
- C. Concatenación: contexto, continuidad y prosodia
- R. espectral: Modelos sinusoidal (I)/ ninguna (II)
- Otras: el sistema II tiene menos unidades y la información de onsets se obtiene mediante análisis sinusoidal.

Participante: UPC

- Unidad base: Semifonema con contexto
- Tipo: Concatenación
- Modelo prosódico: entonación basada en datos (CART)
- C. Objetivo: prosodia, contexto, palabra, acento,...
- C. Concatenación: F0, criterios fonológicos, distancia espectral.
- R. espectral:-
- Otras: Pratt para obtener el pitch.

Participante: UPV-EHU

- Unidad base: Semifonema con contexto
- Tipo: Concatenación
- Modelo prosódico: entonación basada en datos (CART)
- C. Objetivo: 5 fonemas, prosodia, acento,…
- C. Concatenación: F0, energía, distancia espectral.
- R. espectral:-
- Otras: ESPS más algoritmos propios para obtener el pitch.

Participante: URL

- Unidad base: Difonema/Trifonema
- Tipo: Concatenación
- Modelo prosódico: -
- C. Objetivo: -
- C. Concatenación: F0, energía y MFCC.
- R. espectral: MFCC para costes.
- Otras: get_f0 para el pitch con postproceso.

Participante: UVigo

- Unidad base: Semifonema con contexto
- Tipo: Concatenación
- Modelo prosódico: basado en corpus
- C. Objetivo: F0, energía y MFCC
- C. Concatenación: F0, energía y MFCC.
- R. espectral: MFCC.
- Otras: Pratt para el pitch.

Resultados

Sist	Median	MAD	Mean	SD	Samples
I	5	0.0	4.82	0.41	212
C	3	1.0	3.34	0.92	524
E	3	1.0	3.20	0.89	524
B	3	1.0	2.91	0.93	524
H	3	1.0	2.86	0.96	524
F	3	1.0	2.81	0.91	524
D	3	1.0	2.60	0.94	524
G	3	1.0	2.56	0.91	524
A	2	1.0	2.28	0.94	524

Tabla 2. *MOS Naturalidad para todos los oyentes.*

Sist	WER	Samples	Words	S	I	D
C	0.03	188	1233	29	5	9
E	0.05	188	1234	40	4	15
B	0.05	188	1233	47	13	1
H	0.06	188	1234	68	8	3
F	0.08	188	1233	62	29	10
D	0.07	188	1235	61	20	8
G	0.04	188	1234	30	3	12
A	0.08	188	1234	70	4	22

Tabla 7. *WER (substitutions, insertion, deletions) para oyentes nativos*

Sist	Median	MAD	Mean	SD	Samples
I	5	0	4.11	1.17	107
C	3	1	3.25	0.84	107
E	3	1	3.35	0.94	107
B	3	1	3.36	0.93	107
H	3	1	3.29	0.91	107
F	3	1	3.23	0.81	107
D	3	1	3.11	0.89	107
G	3	1	3.07	0.92	107
A	3	1	2.96	0.98	107

Tabla 5. *MOS Similitud para todos los oyentes.*

Características de la evaluación

- Organización: **Nerea Ezeiza**
- Objetivo: Traducción de texto en Castellano a Euskera
- **Entradas:** frases de texto en Castellano
- **Salidas:** traducción en Euskera
- 3 participantes: todas universidades

Descripción del Corpus

- Artículos de divulgación.
- Conjunto de entrenamiento: 58000 frases en Castellano y Euskera.
- Conjunto de validación: 1500 frases en Castellano.
- Medidas de evaluación: BLEU, NIST, WER y PER (TC-STAR)

Participante: AVIVAVOZ

- Tecnología: Phrase-base built with Moses.
- Descripción del sistema:
  - Grow-diagonal-final.
  - Maximum phrase-length: 5.
  - 5-gram language model.
  - 7-gram POS language model on target.
  - Alignment obtained through segmented Basque corpus.
  - 7-gram lemma language model on target.

Participante: PRHLT-UPV

- Tecnología: SITG (Stochastic Inversion Transduction Grammars) with 5 non-terminal symbols.
- Descripción del sistema:
  - 2 reestimation iterations.
  - Log-linear approximation:
    - Direct and inverse translation models (obtained by counting occurrences).
    - Direct and inverse syntactic models.
    - Direct and inverse lexicalized models.
  - Monotonic search.

Participante: IXA-EHU

- Tecnología: combinación de varios sistemas
- Descripción del sistema:
  - Primer sistema:
    - GIZA++ para el alineamiento de palabras y MOSES decoder para la traducción.
    - Modelo de lenguaje basado en 5-gramas.
  - Matrex:
    - Se amplia la tabla de traducción con nuevos phrases
    - Se alinean los sintagmas basándose en los alin. IBM1 y Los sintagmas alineados se incorporan a la tabla de traducción
  - Seg:
    - Se segmenta el texto en Euskera, separando las palabras en diferentes tokens.
  - Combinación de Matrex y Seg:
    - Se amplia la tabla de traducción con nuevos phrases y se segmenta el texto en Euskera, separando las palabras en diferentes tokens.

Resultados

SISTEMA	BLEU	NIST	PER	WER	Dbleu	Dnist	Dper	Dwer	Total
Avivavoz-final	0.0812	3.8985	64.2203	81.6021	1.000	0.746	0.405	0.264	2.414
PRHLT.n5.g2.VII	0.0711	3.6522	65.5574	82.6420	0.000	0.000	0.000	0.000	0.000
IXA.matrex+seg	0.0810	3.9825	62.2534	78.7022	0.980	1.000	1.000	1.000	3.980

Resultados de los sistemas finales

$$D_{bleu} = \frac{BLEUi - MIN}{MAX - MIN} \qquad D_{nist} = \frac{NISTi - MIN}{MAX - MIN}$$

$$D_{WER} = 1 - \frac{WERi - MIN}{MAX - MIN} \qquad D_{PER} = 1 - \frac{PERi - MIN}{MAX - MIN}$$

¡¡Muchas gracias a todas las personas que habéis hecho posible esta Evaluación Albayzín 2008:

organizadores y participantes!!