KALAKA: A TV Broadcast Speech Database for the Evaluation of Language Recognition Systems

Luis J. Rodríguez-Fuentes, Mikel Penagarikano, Germán Bordel, Amparo Varona, Mireia Díez

Software Technologies Working Group (http://gtts.ehu.es)
Department of Electricity and Electronics, University of the Basque Country
Barrio Sarriena s/n, 48940 Leioa, Spain
email: luisjavier.rodriguez@ehu.es

LREC 2010, La Valletta, Malta
May 20, 2010

Contents

1. Introduction
    - Motivation
    - Database features (in brief)
2. Design issues
3. Recording setup
4. Creating the database
    - Classification of recordings
    - Selection of speech segments
    - Automatic extraction of 30-, 10- and 3-second segments
5. Using the database
    - The Albayzin 2008 LRE
    - Developing language recognition technology
6. Conclusions and future work

GTTS
Tecnologías Software

Motivation

- To support the Albayzin 2008 Language Recognition Evaluation, organized by the Spanish Network on Speech Technologies, from May to November 2008.

Motivation

- To support the Albayzin 2008 Language Recognition Evaluation, organized by the Spanish Network on Speech Technologies, from May to November 2008.
- To solve the lack of a multilingual speech database specifically designed for language recognition applications featuring the official languages in Spain as target languages.

Motivation

- To support the Albayzin 2008 Language Recognition Evaluation, organized by the Spanish Network on Speech Technologies, from May to November 2008.
- To solve the lack of a multilingual speech database specifically designed for language recognition applications featuring the official languages in Spain as target languages.
- To build a language recognition module for the backend of an audio indexing and retrieval system dealing with wide-band broadcast news in Spanish and Basque.

Motivation

- To support the Albayzin 2008 Language Recognition Evaluation, organized by the Spanish Network on Speech Technologies, from May to November 2008.
- To solve the lack of a multilingual speech database specifically designed for language recognition applications featuring the official languages in Spain as target languages.
- To build a language recognition module for the backend of an audio indexing and retrieval system dealing with wide-band broadcast news in Spanish and Basque.
- To measure the accuracy that state-of-the-art language recognition systems can attain for the task of recognizing four target languages that have evolved (and continue evolving) in close contact each other.

Motivation

- To support the Albayzin 2008 Language Recognition Evaluation, organized by the Spanish Network on Speech Technologies, from May to November 2008.
- To solve the lack of a multilingual speech database specifically designed for language recognition applications featuring the official languages in Spain as target languages.
- To build a language recognition module for the backend of an audio indexing and retrieval system dealing with wide-band broadcast news in Spanish and Basque.
- To measure the accuracy that state-of-the-art language recognition systems can attain for the task of recognizing four target languages that have evolved (and continue evolving) in close contact each other.
- *May this task be more challenging than expected?*

Database features (in brief)

- Four target languages: Spanish, Catalan, Basque and Galician.

Database features (in brief)

- Four target languages: Spanish, Catalan, Basque and Galician.
- Other (european) languages (to allow open-set tests): French, Portuguese, German and English.

Database features (in brief)

- Four target languages: Spanish, Catalan, Basque and Galician.
- Other (european) languages (to allow open-set tests): French, Portuguese, German and English.
- Speech signals extracted from TV shows, including both planned and spontaneous speech in diverse environment conditions involving a varying number of speakers.

Database features (in brief)

- Four target languages: Spanish, Catalan, Basque and Galician.
- Other (european) languages (to allow open-set tests): French, Portuguese, German and English.
- Speech signals extracted from TV shows, including both planned and spontaneous speech in diverse environment conditions involving a varying number of speakers.
- Size: around 50 hours (3 DVD)
  - Train dataset: 36 hours (9 hours per target language)
  - Development dataset: 7,7 hours (90 minutes per target language + 90 minutes of other languages all together)
  - Evaluation dataset: 7,7 hours (90 minutes per target language + 90 minutes of other languages all together)

Design issues

- Basic design criteria:

  1. Regarding recording setup (devices, connectors, audio conversions, etc.): the same for all the languages
  2. Regarding other sources of variability (environment, speaker, etc.): as much diversity as possible

Design issues

- Basic design criteria:
    1. Regarding recording setup (devices, connectors, audio conversions, etc.): the same for all the languages
    2. Regarding other sources of variability (environment, speaker, etc.): as much diversity as possible
- Cable TV: easy access to audio in different languages

Design issues

- Basic design criteria:
  1. Regarding recording setup (devices, connectors, audio conversions, etc.): the same for all the languages
  2. Regarding other sources of variability (environment, speaker, etc.): as much diversity as possible
- Cable TV: easy access to audio in different languages
- Disjoint subsets of TV shows assigned to train, development and evaluation

Design issues

- Basic design criteria:

  1. Regarding recording setup (devices, connectors, audio conversions, etc.): the same for all the languages
  2. Regarding other sources of variability (environment, speaker, etc.): as much diversity as possible

- Cable TV: easy access to audio in different languages

- Disjoint subsets of TV shows assigned to train, development and evaluation

- Regarding duration:

  - Train dataset: no constraints
  - Development and evaluation datasets: three subsets, containing segments of three nominal durations: 30, 10 and 3 seconds

Recording setup

- Roland Edirol R-09 ultra-light audio recorder

Recording setup

- Roland Edirol R-09 ultra-light audio recorder
- CD quality (16 bit / 44.1 kHz / stereo) recordings

Recording setup

- Roland Edirol R-09 ultra-light audio recorder
- CD quality (16 bit / 44.1 kHz / stereo) recordings
- Audio signals downsampled to 16 kHz, single channel, by means of SoX

Recording setup

- Roland Edirol R-09 ultra-light audio recorder
- CD quality (16 bit / 44.1 kHz / stereo) recordings
- Audio signals downsampled to 16 kHz, single channel, by means of SoX
- Recordings filtered to discard noisy segments

Recording setup

- Roland Edirol R-09 ultra-light audio recorder
- CD quality (16 bit / 44.1 kHz / stereo) recordings
- Audio signals downsampled to 16 kHz, single channel, by means of SoX
- Recordings filtered to discard noisy segments
- Size of recorded materials (138 hours): 3 times the size of speech segments finally used in KALAKA

Recording setup

TV channels and recorded time (in minutes) for each language in KALAKA

Language	***TV Channels***	***Recorded time***
Spanish	TVE1, La 2, La Sexta, Cuatro, Tele5, Antena3, ETB2, TV Canaria Sat, AndalucíaTV, TeleMadrid	1818
Catalan	TVCi	1777
Basque	ETB1	1905
Galician	TVG	1731
German	DWTV	275
French	TV5Monde Europe	320
English	DWTV, BBCWorld	257
Portuguese	RTPi	218

Classification of recordings: target languages

- **Task:** distribute TV shows into three datasets (train, development and evaluation)

Classification of recordings: target languages

- **Task:** distribute TV shows into three datasets (train, development and evaluation)
- **Two basic criteria:**
  - *independence*: a given TV show is always posted to the same dataset
  - *diversity*: similar proportions of show types in all datasets

Classification of recordings: target languages

- **Task:** distribute TV shows into three datasets (train, development and evaluation)
- **Two basic criteria:**
  - *independence*: a given TV show is always posted to the same dataset
  - *diversity*: similar proportions of show types in all datasets
- TV show types: (1) debates and interviews; (2) talk-shows; (3) news; (4) sports; (5) entertaining (contests, reality shows, etc.); and (6) documentaries

Classification of recordings: target languages

- **Task:** distribute TV shows into three datasets (train, development and evaluation)
- **Two basic criteria:**
  - *independence*: a given TV show is always posted to the same dataset
  - *diversity*: similar proportions of show types in all datasets
- TV show types: (1) debates and interviews; (2) talk-shows; (3) news; (4) sports; (5) entertaining (contests, reality shows, etc.); and (6) documentaries
- Most debates and interviews posted to the train dataset

Classification of recordings: target languages

Recorded time, absolute (minutes) and relative (%), of the six types of TV shows for the target languages.

	Spanish	***Catalan***	***Basque***	***Galician***
Debates	495 - 27.23	499 - 28.08	631 - 33.12	515 - 29.75
Talk-shows	500 - 27.50	428 - 24.09	498 - 26.14	642 - 37.09
News	353 - 19.42	336 - 18.91	341 - 17.90	405 - 23.40
Sports	126 - 6.93	120 - 6.75	120 - 6.30	17 - 0.98
Entertaining	230 - 12.65	249 - 14.01	153 - 8.03	83 - 4.79
Documentaries	114 - 6.27	145 - 8.16	162 - 8.50	69 - 3.99
Total	1818 - 100.00	1777 - 100.00	1905 - 100.00	1731 - 100.00

Classification of recordings: non-target languages

- TV shows corresponding to non-target languages posted to development and evaluation.

Classification of recordings: non-target languages

- TV shows corresponding to non-target languages posted to development and evaluation.
- Proportions made deliberately different for development and evaluation, to avoid tuning systems to reject specific languages.

Classification of recordings: non-target languages

- TV shows corresponding to non-target languages posted to development and evaluation.
- Proportions made deliberately different for development and evaluation, to avoid tuning systems to reject specific languages.
- Proportion of French and Portuguese twice the proportion of German and English.

Classification of recordings: non-target languages

- TV shows corresponding to non-target languages posted to development and evaluation.
- Proportions made deliberately different for development and evaluation, to avoid tuning systems to reject specific languages.
- Proportion of French and Portuguese twice the proportion of German and English.

Planned distribution of data (%) for non-target languages

	Dev	***Eval***	***Total***
German	0.00	16.67	16.67
French	29.17	4.16	33.33
English	16.67	0.00	16.67
Portuguese	4.16	29.17	33.33
Total	50.00	50.00	100.00

Selection of speech segments

- **Task:** to extract speech segments from recorded materials

Selection of speech segments

- **Task:** to extract speech segments from recorded materials
- **Criteria:**
  - high SNR (clean speech or low-level background noise)
  - no speech overlaps
  - single language

Selection of speech segments

- **Task:** to extract speech segments from recorded materials
- **Criteria:**
  - high SNR (clean speech or low-level background noise)
  - no speech overlaps
  - single language
- **Tools:** *Wavesurfer* and *CoolEdit*

Selection of speech segments

- **Task:** to extract speech segments from recorded materials
- **Criteria:**
  - high SNR (clean speech or low-level background noise)
  - no speech overlaps
  - single language
- **Tools:** *Wavesurfer* and *CoolEdit*
- **Result:** speech segments of indefinite length, spoken by one or more speakers in a single language

Selection of speech segments

- **Task:** to extract speech segments from recorded materials
- **Criteria:**
  - high SNR (clean speech or low-level background noise)
  - no speech overlaps
  - single language
- **Tools:** *Wavesurfer* and *CoolEdit*
- **Result:** speech segments of indefinite length, spoken by one or more speakers in a single language
- No further processing applied to segments posted to the train dataset

Selection of speech segments

- **Task:** to extract speech segments from recorded materials
- **Criteria:**
  - high SNR (clean speech or low-level background noise)
  - no speech overlaps
  - single language
- **Tools:** *Wavesurfer* and *CoolEdit*
- **Result:** speech segments of indefinite length, spoken by one or more speakers in a single language
- No further processing applied to segments posted to the train dataset

Segments posted to the train dataset in KALAKA.

	Spanish	***Catalan***	***Basque***	***Galician***	***All***
# segments	282	278	342	401	1303
Duration (min)	529	538	531	532	2130

Automatic extraction of 30-, 10- and 3-second segments

- Speech segments posted to development and evaluation, taken as source to extract segments of fixed nominal duration: 30, 10 and 3 seconds

Automatic extraction of 30-, 10- and 3-second segments

- Speech segments posted to development and evaluation, taken as source to extract segments of fixed nominal duration: 30, 10 and 3 seconds
- **Criteria:**
  1. Each segment enclosed by a certain amount of silence

Automatic extraction of 30-, 10- and 3-second segments

- Speech segments posted to development and evaluation, taken as source to extract segments of fixed nominal duration: 30, 10 and 3 seconds
- **Criteria:**
  1. Each segment enclosed by a certain amount of silence
  2. A 30-second segment is validated if and only if it contains a valid 10-second segment. Similarly, a 10-second segment is validated if and only if it contains a 3-second segment

Automatic extraction of 30-, 10- and 3-second segments

- Speech segments posted to development and evaluation, taken as source to extract segments of fixed nominal duration: 30, 10 and 3 seconds
- **Criteria:**
  1. Each segment enclosed by a certain amount of silence
  2. A 30-second segment is validated if and only if it contains a valid 10-second segment. Similarly, a 10-second segment is validated if and only if it contains a 3-second segment
  3. Segments can be slightly longer than their nominal duration

Automatic extraction of 30-, 10- and 3-second segments

- Speech segments posted to development and evaluation, taken as source to extract segments of fixed nominal duration: 30, 10 and 3 seconds
- **Criteria:**
  1. Each segment enclosed by a certain amount of silence
  2. A 30-second segment is validated if and only if it contains a valid 10-second segment. Similarly, a 10-second segment is validated if and only if it contains a 3-second segment
  3. Segments can be slightly longer than their nominal duration
- Performance differences measured on these subsets due (we expect) to the varying amount of available speech

Automatic extraction of 30-, 10- and 3-second segments

- Speech segments posted to development and evaluation, taken as source to extract segments of fixed nominal duration: 30, 10 and 3 seconds
- **Criteria:**
  1. Each segment enclosed by a certain amount of silence
  2. A 30-second segment is validated if and only if it contains a valid 10-second segment. Similarly, a 10-second segment is validated if and only if it contains a 3-second segment
  3. Segments can be slightly longer than their nominal duration
- Performance differences measured on these subsets due (we expect) to the varying amount of available speech
- Single-pass greedy algorithm, retrieving 65% of the input speech

Automatic extraction of 30-, 10- and 3-second segments

- Speech segments posted to development and evaluation, taken as source to extract segments of fixed nominal duration: 30, 10 and 3 seconds
- **Criteria:**
  1. Each segment enclosed by a certain amount of silence
  2. A 30-second segment is validated if and only if it contains a valid 10-second segment. Similarly, a 10-second segment is validated if and only if it contains a 3-second segment
  3. Segments can be slightly longer than their nominal duration
- Performance differences measured on these subsets due (we expect) to the varying amount of available speech
- Single-pass greedy algorithm, retrieving 65% of the input speech
- Result (development and evaluation):
  - Total: 1800 segments
  - 600 segments per duration
  - 120 segments per target language and duration
  - 120 segments of non-target languages all together per duration (different distributions for development and evaluation)

The Albayzin 2008 LRE

- **Task:** independent language verification trials for a set of 4 target languages: Basque, Catalan, Galician and Spanish

The Albayzin 2008 LRE

- **Task:** independent language verification trials for a set of 4 target languages: Basque, Catalan, Galician and Spanish
- **Test conditions:**
  - 30-, 10- and 3-second test segments
  - Restricted vs. free system development
  - Closed-set vs. open-set evaluation

The Albayzin 2008 LRE

- **Task:** independent language verification trials for a set of 4 target languages: Basque, Catalan, Galician and Spanish
- **Test conditions:**
  - 30-, 10- and 3-second test segments
  - Restricted vs. free system development
  - Closed-set vs. open-set evaluation
- **Award:** system yielding the least cost in the restricted-development closed-set test condition on the subset of 30-second speech segments

The Albayzin 2008 LRE

- **Task:** independent language verification trials for a set of 4 target languages: Basque, Catalan, Galician and Spanish
- **Test conditions:**
  - 30-, 10- and 3-second test segments
  - Restricted vs. free system development
  - Closed-set vs. open-set evaluation
- **Award:** system yielding the least cost in the restricted-development closed-set test condition on the subset of 30-second speech segments
- **Average performance**
  - Two sites applying state-of-the-art language recognition systems

The Albayzin 2008 LRE

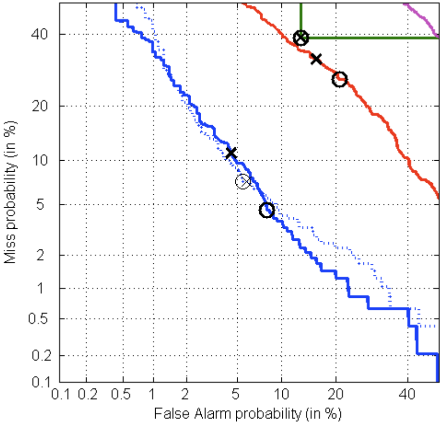
- **Task:** independent language verification trials for a set of 4 target languages: Basque, Catalan, Galician and Spanish
- **Test conditions:**
  - 30-, 10- and 3-second test segments
  - Restricted vs. free system development
  - Closed-set vs. open-set evaluation
- **Award:** system yielding the least cost in the restricted-development closed-set test condition on the subset of 30-second speech segments
- **Average performance**
  - Two sites applying state-of-the-art language recognition systems
  - Free-development, closed-set, 30-second segments: 5% EER

The Albayzin 2008 LRE

- **Task:** independent language verification trials for a set of 4 target languages: Basque, Catalan, Galician and Spanish
- **Test conditions:**
  - 30-, 10- and 3-second test segments
  - Restricted vs. free system development
  - Closed-set vs. open-set evaluation
- **Award:** system yielding the least cost in the restricted-development closed-set test condition on the subset of 30-second speech segments
- **Average performance**
  - Two sites applying state-of-the-art language recognition systems
  - Free-development, closed-set, 30-second segments: 5% EER
  - Free-development, open-set, 30-second segments: 9% EER

The Albayzin 2008 LRE

Pooled DET curves of systems in the restricted-development closed-set test condition on 30-second speech segments.



GTTS Language Recognition System - Features

- Train and development materials: KALAKA

GTTS Language Recognition System - Features

- Train and development materials: KALAKA
- Hierarchical fusion of one acoustic subsystem (GMM-SVM) and 6 phonotactic subsytems (3 Phone-LM + 3 Phone-SVM)

GTTS Language Recognition System - Features

- Train and development materials: KALAKA
- Hierarchical fusion of one acoustic subsystem (GMM-SVM) and 6 phonotactic subsytems (3 Phone-LM + 3 Phone-SVM)
- **GMM-SVM:** SVM classifier on the vector space defined by the means of a Gaussian Mixture Model

GTTS Language Recognition System - Features

- Train and development materials: KALAKA
- Hierarchical fusion of one acoustic subsystem (GMM-SVM) and 6 phonotactic subsytems (3 Phone-LM + 3 Phone-SVM)
- **GMM-SVM:** SVM classifier on the vector space defined by the means of a Gaussian Mixture Model
- **Phonotactic subsystems:**
    - BUT TRAPS/NN phone decoders for Czech, Hungarian and Russian

GTTS Language Recognition System - Features

- Train and development materials: KALAKA
- Hierarchical fusion of one acoustic subsystem (GMM-SVM) and 6 phonotactic subsytems (3 Phone-LM + 3 Phone-SVM)
- **GMM-SVM:** SVM classifier on the vector space defined by the means of a Gaussian Mixture Model
- **Phonotactic subsystems:**
  - BUT TRAPS/NN phone decoders for Czech, Hungarian and Russian
  - Phone decodings computed on signals downsampled to 8 kHz

GTTS Language Recognition System - Features

- Train and development materials: KALAKA
- Hierarchical fusion of one acoustic subsystem (GMM-SVM) and 6 phonotactic subsytems (3 Phone-LM + 3 Phone-SVM)
- **GMM-SVM:** SVM classifier on the vector space defined by the means of a Gaussian Mixture Model
- **Phonotactic subsystems:**
    - BUT TRAPS/NN phone decoders for Czech, Hungarian and Russian
    - Phone decodings computed on signals downsampled to 8 kHz
    - Sequence modeling approaches:
        - **Phone-LM:** 4-grams with Witten-Bell smoothing
        - **Phone-SVM:** SVM (linear kernel) on bag-of-ngrams (up to 3-grams)

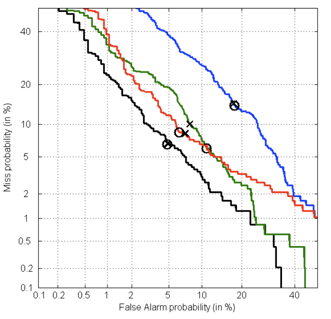
GTTS Language Recognition System - Results

KALAKA: closed-set test condition, 30-second speech segments.

Left: C_{avg} of single and fused language recognition systems.

Right: pooled DET curves of GMM-SVM (blue), fused Phone-LM (green), fused Phone-SVM (red) and the system fusing all of them (black).

		C_{avg}
Single	GMM-SVM	0.1611
	PHONE (CH) - LM	0.1545
	PHONE (HU) - LM	0.1427
	PHONE (RU) - LM	0.1305
	PHONE (CH) - SVM	0.0940
	PHONE (HU) - SVM	0.1017
	PHONE (RU) - SVM	0.1215
Fused	PHONE - LM	0.0892
	PHONE - SVM	0.0774
	PHONE	0.0691
	ALL	**0.0576**



Conclusions

- KALAKA, a database containing speech from TV broadcasts, allows to develop language recognition systems for the official languages in Spain: Basque, Catalan, Galician and Spanish.

Conclusions

- KALAKA, a database containing speech from TV broadcasts, allows to develop language recognition systems for the official languages in Spain: Basque, Catalan, Galician and Spanish.
- Results using state-of-the-art technology provide evidence of the difficulty of various tasks defined on KALAKA.

Conclusions

- KALAKA, a database containing speech from TV broadcasts, allows to develop language recognition systems for the official languages in Spain: Basque, Catalan, Galician and Spanish.
- Results using state-of-the-art technology provide evidence of the difficulty of various tasks defined on KALAKA.
- KALAKA can be challenging enough to support further developments in language recognition technology.

Future work

Actually, current work: $\boxed{\text{KALAKA-2}}$

Future work

Actually, current work: $\boxed{\text{KALAKA-2}}$

- An extended version of KALAKA, adding Portuguese and English as target languages, renewing the set of unknown languages and including a new test condition for noisy speech.

Future work

Actually, current work: $\boxed{\text{KALAKA-2}}$

- An extended version of KALAKA, adding Portuguese and English as target languages, renewing the set of unknown languages and including a new test condition for noisy speech.

- Support for the Albayzin 2010 LRE: June to October 2010, results presented at FALA 2010, to be held in Vigo (Spain) in November 2010.

Future work

Actually, current work: KALAKA-2

- An extended version of KALAKA, adding Portuguese and English as target languages, renewing the set of unknown languages and including a new test condition for noisy speech.

- Support for the Albayzin 2010 LRE: June to October 2010, results presented at FALA 2010, to be held in Vigo (Spain) in November 2010.

- Registration now open at http://fala2010.uvigo.es