KALAKA-2
A TV Broadcast Speech Database for the Recognition of Iberian Languages in Clean and Noisy Environments

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

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 2012, Istanbul, Turkey
May 23, 2012

GTTS Tecnologías Software

Contents

GTTS Tecnologías Software

Motivation

- To support the **Albayzin 2010 Language Recognition Evaluation**, organized by the Spanish Network on Speech Technologies, from May to November 2010 (**second edition**).

Motivation

- To support the **Albayzin 2010 Language Recognition Evaluation**, organized by the Spanish Network on Speech Technologies, from May to November 2010 (**second edition**).
- To provide a multilingual speech database specifically designed for language recognition applications featuring Iberian languages as target languages (**including Portuguese and English**).

Motivation

- To support the **Albayzin 2010 Language Recognition Evaluation**, organized by the Spanish Network on Speech Technologies, from May to November 2010 (**second edition**).
- To provide a multilingual speech database specifically designed for language recognition applications featuring Iberian languages as target languages (**including Portuguese and English**).
- To measure the performance of state-of-the-art language recognition systems when dealing with **noisy/overlapped speech**, and compare it to the performance on clean speech.

Database features (in brief)

- Six target languages: Basque, Catalan, English, Galician, Portuguese and Spanish.

Database features (in brief)

- Six target languages: Basque, Catalan, English, Galician, Portuguese and Spanish.
- Out-Of-Set (OOS) languages, to allow open-set tests: Arabic, French, German and Romanian.

Database features (in brief)

- Six target languages: Basque, Catalan, English, Galician, Portuguese and Spanish.
- Out-Of-Set (OOS) languages, to allow open-set tests: Arabic, French, German and Romanian.
- Speech signals extracted from TV shows, including planned and spontaneous speech involving a varying number of speakers.

Database features (in brief)

- Six target languages: Basque, Catalan, English, Galician, Portuguese and Spanish.
- Out-Of-Set (OOS) languages, to allow open-set tests: Arabic, French, German and Romanian.
- Speech signals extracted from TV shows, including planned and spontaneous speech involving a varying number of speakers.
- Two types of speech signals: **clean** (mostly studio conditions) and **noisy** (noise, music or speech in the background, or overlapped speech)

Database features (in brief)

- Six target languages: Basque, Catalan, English, Galician, Portuguese and Spanish.
- Out-Of-Set (OOS) languages, to allow open-set tests: Arabic, French, German and Romanian.
- Speech signals extracted from TV shows, including planned and spontaneous speech involving a varying number of speakers.
- Two types of speech signals: **clean** (mostly studio conditions) and **noisy** (noise, music or speech in the background, or overlapped speech)
- Database size: around 125 hours (**5 DVD, by request to authors**)
  - **Training** dataset: 82 hours (more than 13 hours per target language, 80% clean + 20% noisy)
  - **Development** dataset: 21,5 hours (4950 segments, 3 nominal durations, target and OOS languages, 70% clean + 30% noisy)
  - **Evaluation** dataset: 21,5 hours (4992 segments, 3 nominal durations, target and OOS languages, 67% clean + 33% noisy)

GTTS Tecnologías Software

Design issues

- **Basic design criteria:**
  - Single recording setup (devices, connectors, audio conversions, etc.)
  - All the materials classified into: (1) clean or (2) noisy/overlapped
  - Other sources of variability (speakers, etc.): as much diversity as possible

Design issues

- **Basic design criteria:**
  - Single recording setup (devices, connectors, audio conversions, etc.)
  - All the materials classified into: (1) clean or (2) noisy/overlapped
  - Other sources of variability (speakers, etc.): as much diversity as possible
- **KALAKA-2 is a major update of KALAKA:**
  - Two new target languages: Portuguese and English
  - KALAKA materials fully recycled:

    KALAKA train + dev $\rightarrow$ KALAKA-2 train
    KALAKA eval $\rightarrow$ KALAKA-2 dev
  - New recordings (specially for Portuguese, English and OOS languages)
  - Disjoint subsets of TV shows assigned to train, dev and eval
  - **Evaluation dataset entirely built on new recordings**

GTTS Tecnologías Software

Design issues

- **Basic design criteria:**
  - Single recording setup (devices, connectors, audio conversions, etc.)
  - All the materials classified into: (1) clean or (2) noisy/overlapped
  - Other sources of variability (speakers, etc.): as much diversity as possible
- **KALAKA-2 is a major update of KALAKA:**
  - Two new target languages: Portuguese and English
  - KALAKA materials fully recycled:

    KALAKA train + dev $\rightarrow$ KALAKA-2 train
    KALAKA eval $\rightarrow$ KALAKA-2 dev
  - New recordings (specially for Portuguese, English and OOS languages)
  - Disjoint subsets of TV shows assigned to train, dev and eval
  - **Evaluation dataset entirely built on new recordings**
- **Segment duration:**
  - train: no constraints
  - dev and eval: three nominal durations of 30, 10 and 3 seconds

GTTS Tecnologías Software

Recording setup (I)

- Cable TV: easy access to audio in different languages
- 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
- Three recording times:
  - October-November 2008 (Arabic, Romanian and English)
  - April-May 2010 (Arabic, German, French, Romanian, English and Portuguese)
  - August-September 2010 (Basque, Catalan, Galician and Spanish)
- Recorded time: 257 hours (more than 2 times the size of KALAKA-2)

GTTS Tecnologías Software

Recording setup (II)

TV channels and recorded time (in minutes) for each language in KALAKA-2

Language	TV Channels	Recorded time
Basque	ETB1, ETBSat	1996
Catalan	TVCi	1842
English	DWTV, BBCWorld, CNN, Bloomberg	2705
Galician	TVG	2240
Portuguese	RTPi	2608
Spanish	TVE1, La 2, La Sexta, Cuatro, Tele5, Antena3, ETB2, TV Canaria Sat, AndalucíaTV, TeleMadrid, ExtremaduraTV, CNNPlus	2090
Arabic	Al Jazeera	497
French	TV5Monde Europe	499
German	DWTV	431
Romanian	PROTV	552

GTTS Tecnologías Software

Classification of recordings

- **Task:** distribute TV shows into training, 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
- Different distributions of OOS languages for development and evaluation, to avoid tuning systems to reject specific OOS languages.

Selection of speech segments (I)

- **Task:** to extract speech segments from the recorded materials, by listening and looking at audio signals.
- **Criteria:**
  - Multiple speakers allowed
  - Single (nominal) language
  - Clean/Noisy classification relaxed: *mostly clean* and *mostly noisy* segments
  - **Discarded:** (1) narrow-band (telephone-channel) speech and (2) fragments with multiple languages (even in the background)
  - **Exception:** two or more OOS languages may appear in the same segment
- **Tools:** *Wavesurfer* and *CoolEdit*
- **Results:**
  - *clean speech*: segments of any length greater than 30 seconds
  - *noisy/overlapped speech*: segments of length between 30 and 35 seconds

GTTS Tecnologías Software

Selection of speech segments (II) - Training dataset

- No further processing was applied to segments posted to training.
- Training data **ONLY** for target languages.
- More than 10 hours of clean speech and more than 2 hours of noisy speech per target language.

Distribution of training segments per target language in KALAKA-2, for clean and noisy speech: number of segments (#) and total duration (T, in minutes).

	Clean speech		**Noisy speech**	
	#	**T (minutes)**	**#**	**T (minutes)**
Basque	406	644	112	135
Catalan	341	687	107	131
English	249	731	136	152
Galician	464	644	125	134
Portuguese	387	665	160	197
Spanish	342	625	133	222

GTTS Tecnologías Software

Automatic extraction of 30-, 10- and 3-second segments (I)

- Segments of fixed nominal duration (30, 10 and 3 seconds) extracted from clean-speech fragments posted to dev and eval.
- **Single-pass greedy algorithm:**
  - Segments enclosed by a certain amount of silence.
  - 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.
  - Segments can be slightly longer than their nominal duration.

Automatic extraction of 30-, 10- and 3-second segments (I)

- Segments of fixed nominal duration (30, 10 and 3 seconds) extracted from clean-speech fragments posted to dev and eval.
- **Single-pass greedy algorithm:**
  - Segments enclosed by a certain amount of silence.
  - 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.
  - Segments can be slightly longer than their nominal duration.
- Noisy-speech fragments of 30-35 seconds stored as 30-second segments.
- Greedy algorithm applied to extract 10- and 3-second noisy segments.

Automatic extraction of 30-, 10- and 3-second segments (I)

- Segments of fixed nominal duration (30, 10 and 3 seconds) extracted from clean-speech fragments posted to dev and eval.
- **Single-pass greedy algorithm:**
  - Segments enclosed by a certain amount of silence.
  - 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.
  - Segments can be slightly longer than their nominal duration.
- Noisy-speech fragments of 30-35 seconds stored as 30-second segments.
- Greedy algorithm applied to extract 10- and 3-second noisy segments.
- **Development and evaluation datasets:**
  - Same size and characteristics, except for the distribution of OOS languages and the proportion of clean and noisy speech.
  - At least 150 segments per target language and nominal duration.
  - Around 450 OOS segments per nominal duration.

GTTS Tecnologías Software

Automatic extraction of 30-, 10- and 3-second segments (II)

Distribution of segments per language (the same for each nominal duration) in the development and evaluation datasets of KALAKA-2.

		Devel		Eval	
		clean	**noisy**	**clean**	**noisy**
Target languages	**Basque**	146	29	130	74
	Catalan	120	47	149	55
	English	133	60	135	69
	Galician	137	60	121	83
	Portuguese	164	77	146	58
	Spanish	136	83	125	79
OOS languages	**Arabic**	100	25	115	22
	French	120	32	70	34
	German	108	73	13	32
	Romanian	0	0	111	43

GTTS Tecnologías Software

The Albayzin 2010 LRE: conditions

- **Task:** deciding by computational means whether or not a target language was spoken in a test utterance.
  - Trial = speech segment + target language
  - Required system output: hard decision + likelihood score
  - Performance measured by presenting a set of trials and comparing system decisions with the ground truth.

The Albayzin 2010 LRE: conditions

- **Task:** deciding by computational means whether or not a target language was spoken in a test utterance.
  - Trial = speech segment + target language
  - Required system output: hard decision + likelihood score
  - Performance measured by presenting a set of trials and comparing system decisions with the ground truth.
- **Test conditions:**
  - clean-speech vs. noisy-speech
  - closed-set vs. open-set evaluation
  - 30-, 10- and 3-second test segments

The Albayzin 2010 LRE: conditions

- **Task:** deciding by computational means whether or not a target language was spoken in a test utterance.
  - Trial = speech segment + target language
  - Required system output: hard decision + likelihood score
  - Performance measured by presenting a set of trials and comparing system decisions with the ground truth.
- **Test conditions:**
  - clean-speech vs. noisy-speech
  - closed-set vs. open-set evaluation
  - 30-, 10- and 3-second test segments
- **Primary performance measure:** *average cost* C_{avg}
  Combination of miss and false alarm error rates, pooled across target languages, according to language priors (P_{target}, $P_{non-target}$ and P_{OOS}) and application dependent costs (C_{miss} and C_{fa}).
- **DET curves:** to compare the global performance of two systems.

GTTS Tecnologías Software

The Albayzin 2010 LRE: summary of results (I)

- State-of-the-art language recognition systems

The Albayzin 2010 LRE: summary of results (I)

- State-of-the-art language recognition systems
- **Clean-speech, closed-set, 30-second segments:** $C_{avg} \times 100 = 1.81$
  - Performance comparable to that obtained in similar tasks of NIST LRE.
  - Much better than in Albayzin 2008 LRE ($C_{avg} \times 100 = 5.52$): technology improvements, more training data and less confusable target languages (Portuguese and English).

The Albayzin 2010 LRE: summary of results (I)

- State-of-the-art language recognition systems
- **Clean-speech, closed-set, 30-second segments:** $C_{avg} \times 100 = 1.81$
  - Performance comparable to that obtained in similar tasks of NIST LRE.
  - Much better than in Albayzin 2008 LRE ($C_{avg} \times 100 = 5.52$): technology improvements, more training data and less confusable target languages (Portuguese and English).
- **Dependence on nominal duration:** $C_{avg}(3s) \approx 2 \cdot C_{avg}(10s)$ and $C_{avg}(10s) \approx 2 \cdot C_{avg}(30s)$

The Albayzin 2010 LRE: summary of results (I)

- State-of-the-art language recognition systems
- **Clean-speech, closed-set, 30-second segments:** $C_{avg} \times 100 = 1.81$
  - Performance comparable to that obtained in similar tasks of NIST LRE.
  - Much better than in Albayzin 2008 LRE ($C_{avg} \times 100 = 5.52$): technology improvements, more training data and less confusable target languages (Portuguese and English).
- **Dependence on nominal duration:** $C_{avg}(3s) \approx 2 \cdot C_{avg}(10s)$ and $C_{avg}(10s) \approx 2 \cdot C_{avg}(30s)$
- Clean-speech, **open-set**, 30-second segments: $C_{avg} \times 100 = 2.96$
  - 63.5% cost increase (smaller for 10- and 3-second segments)
  - False alarms related to OOS languages

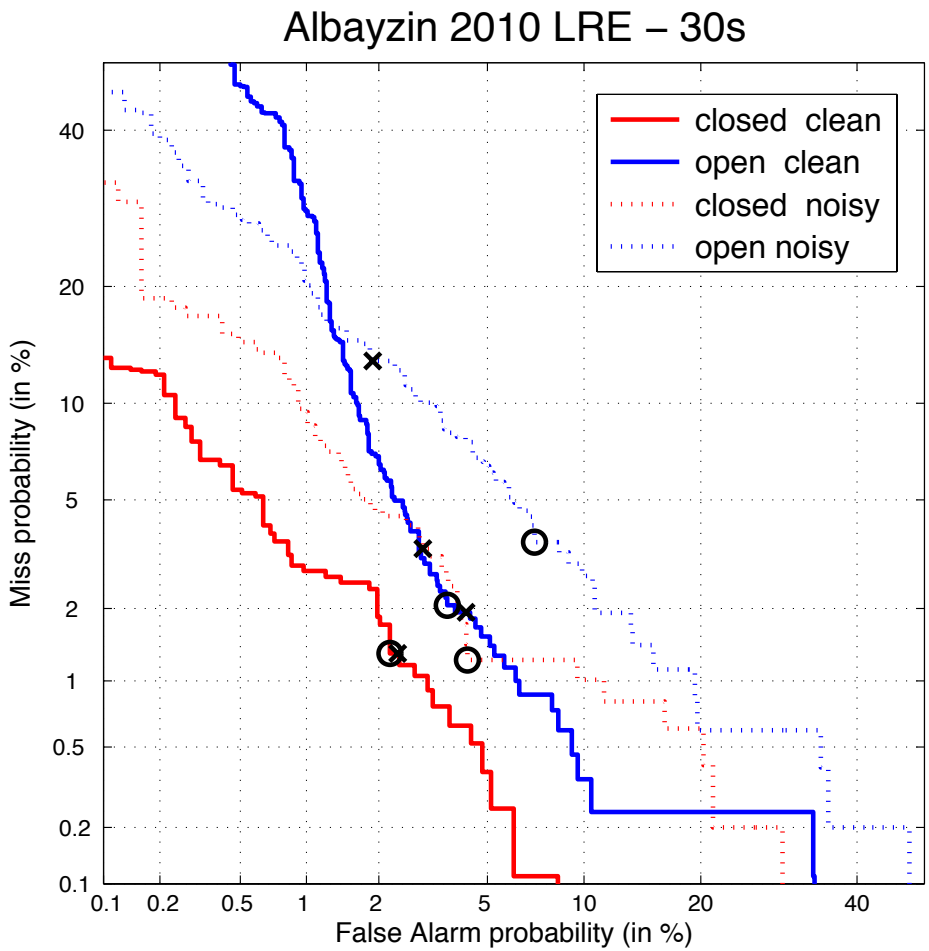
The Albayzin 2010 LRE: summary of results (I)

- State-of-the-art language recognition systems
- **Clean-speech, closed-set, 30-second segments:** $C_{avg} \times 100 = 1.81$
  - Performance comparable to that obtained in similar tasks of NIST LRE.
  - Much better than in Albayzin 2008 LRE ($C_{avg} \times 100 = 5.52$): technology improvements, more training data and less confusable target languages (Portuguese and English).
- **Dependence on nominal duration:** $C_{avg}(3s) \approx 2 \cdot C_{avg}(10s)$ and $C_{avg}(10s) \approx 2 \cdot C_{avg}(30s)$
- Clean-speech, **open-set**, 30-second segments: $C_{avg} \times 100 = 2.96$
  - 63.5% cost increase (smaller for 10- and 3-second segments)
  - False alarms related to OOS languages
- Dealing with **noisy speech**: cost increase ranging from 40% to 80%
- **Noisy-speech, open-set, 3-second segments:** $C_{avg} \times 100 = 15.51$

GTTS Tecnologías Software

The Albayzin 2010 LRE: summary of results (II)

Best primary systems in the 30s tracks of Albayzin 2010 LRE



GTTS Language Recognition System - Features

- System built based **exclusively** on KALAKA-2...

GTTS Language Recognition System - Features

- System built based **exclusively** on KALAKA-2...
  ...except for the phone decoders.

GTTS Language Recognition System - Features

- System built based **exclusively** on KALAKA-2...
  ...except for the phone decoders.
- Same system submitted to NIST 2011 LRE (**with very competitive performance**).
- Discriminative fusion of two acoustic and three phonotactic subsystems:
  - LE-GMM and generative iVectors
  - Phone-Lattice-SVM using BUT decoders for Czech, Hungarian and Russian

GTTS Language Recognition System - Features

- System built based **exclusively** on KALAKA-2...
  ...except for the phone decoders.
- Same system submitted to NIST 2011 LRE (**with very competitive performance**).
- Discriminative fusion of two acoustic and three phonotactic subsystems:
  - LE-GMM and generative iVectors
  - Phone-Lattice-SVM using BUT decoders for Czech, Hungarian and Russian
- Two sets of models, estimated on the training dataset, using:
  - clean speech segments (for the clean-speech condition)
  - the whole training dataset (for the noisy-speech condition)

GTTS Language Recognition System - Features

- System built based **exclusively** on KALAKA-2...
  ...except for the phone decoders.
- Same system submitted to NIST 2011 LRE (**with very competitive performance**).
- Discriminative fusion of two acoustic and three phonotactic subsystems:
  - LE-GMM and generative iVectors
  - Phone-Lattice-SVM using BUT decoders for Czech, Hungarian and Russian
- Two sets of models, estimated on the training dataset, using:
  - clean speech segments (for the clean-speech condition)
  - the whole training dataset (for the noisy-speech condition)
- Generative Gaussian backend and discriminative fusion models estimated on development data, by means of the FoCal toolkit, using:
  - segments containing target languages (for the closed-set condition)
  - all the development segments (for the open-set condition)

GTTS Tecnologías Software

GTTS Language Recognition System - Features

- System built based **exclusively** on KALAKA-2...
  ...except for the phone decoders.
- Same system submitted to NIST 2011 LRE (**with very competitive performance**).
- Discriminative fusion of two acoustic and three phonotactic subsystems:
  - LE-GMM and generative iVectors
  - Phone-Lattice-SVM using BUT decoders for Czech, Hungarian and Russian
- Two sets of models, estimated on the training dataset, using:
  - clean speech segments (for the clean-speech condition)
  - the whole training dataset (for the noisy-speech condition)
- Generative Gaussian backend and discriminative fusion models estimated on development data, by means of the FoCal toolkit, using:
  - segments containing target languages (for the closed-set condition)
  - all the development segments (for the open-set condition)

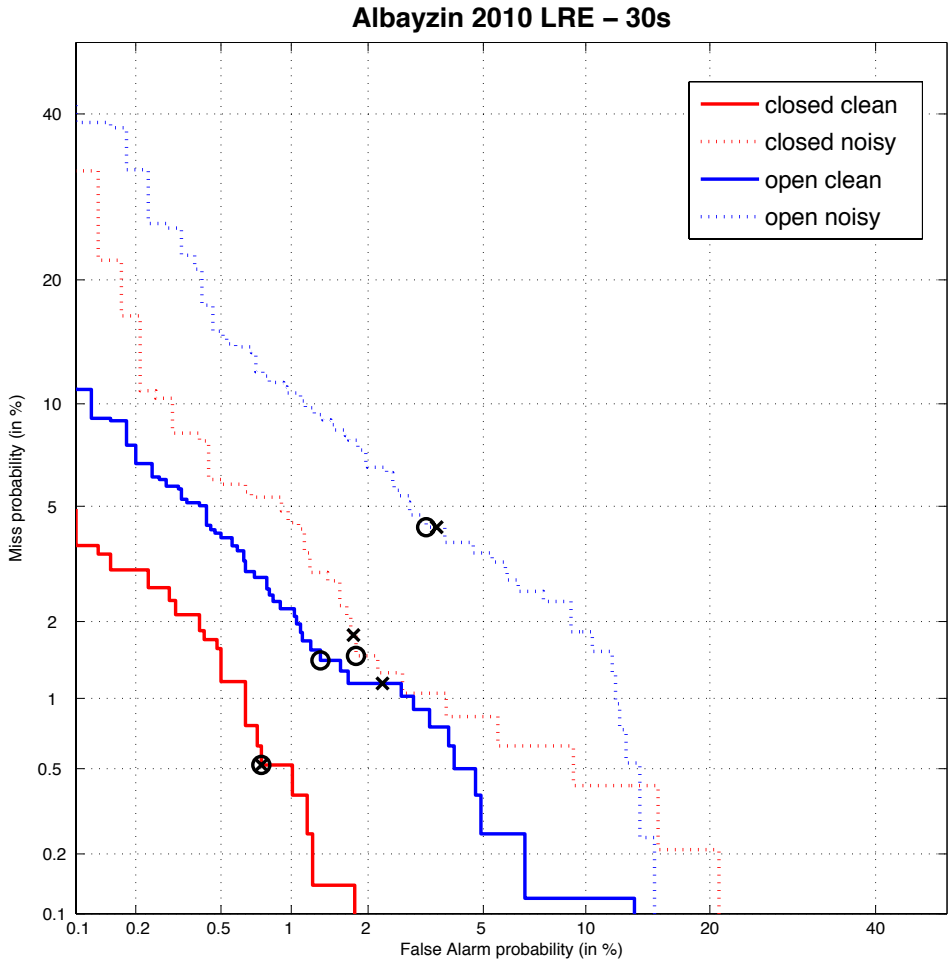
See the paper for more details and references

GTTS Tecnologías Software

GTTS Language Recognition System - Results

C_{avg} performance

	30s	10s	3s
CC	**0.0063**	0.0263	0.0888
OC	0.0171	0.0437	0.1094
CN	0.0177	0.0599	**0.1476**
ON	0.0390	0.0867	**0.1740**



Conclusions

- KALAKA-2: a database containing clean and noisy/overlapped speech from wide-band TV broadcasts is available for your use.

Conclusions

- KALAKA-2: a database containing clean and noisy/overlapped speech from wide-band TV broadcasts is available for your use.
- It allows to develop and evaluate language recognition systems for Iberian languages + English.

Conclusions

- KALAKA-2: a database containing clean and noisy/overlapped speech from wide-band TV broadcasts is available for your use.
- It allows to develop and evaluate language recognition systems for Iberian languages + English.
- Average costs using state-of-the-art technology are high enough to allow further (statistically significant) improvements in all the tracks, except for the easiest one (closed-set, clean-speech, 30s segments).

GTTS Tecnologías Software

Conclusions

- KALAKA-2: a database containing clean and noisy/overlapped speech from wide-band TV broadcasts is available for your use.
- It allows to develop and evaluate language recognition systems for Iberian languages + English.
- Average costs using state-of-the-art technology are high enough to allow further (statistically significant) improvements in all the tracks, except for the easiest one (closed-set, clean-speech, 30s segments).
- Currently, KALAKA-2 is distributed in 5 DVD after direct request to authors.

GTTS Tecnologías Software

Conclusions

- KALAKA-2: a database containing clean and noisy/overlapped speech from wide-band TV broadcasts is available for your use.
- It allows to develop and evaluate language recognition systems for Iberian languages + English.
- Average costs using state-of-the-art technology are high enough to allow further (statistically significant) improvements in all the tracks, except for the easiest one (closed-set, clean-speech, 30s segments).
- Currently, KALAKA-2 is distributed in 5 DVD after direct request to authors.
- We have contacted ELRA to manage licensing issues with the TV broadcast providers.

GTTS Tecnologías Software

Future work

Actually, current work: **KALAKA-3**

Future work

Actually, current work: **KALAKA-3**

- Support for the **Albayzin 2012 Language Recognition Evaluation**: June to October 2012, results presented at **IberSpeech 2012**, to be held in Madrid (Spain) in November 2012.

GTTS Tecnologías Software

Future work

Actually, current work: **KALAKA-3**

- Support for the **Albayzin 2012 Language Recognition Evaluation**: June to October 2012, results presented at **IberSpeech 2012**, to be held in Madrid (Spain) in November 2012.
- Besides including all the materials of KALAKA-2, which will be reused for training,

Future work

Actually, current work: **KALAKA-3**

- Support for the **Albayzin 2012 Language Recognition Evaluation**: June to October 2012, results presented at **IberSpeech 2012**, to be held in Madrid (Spain) in November 2012.
- Besides including all the materials of KALAKA-2, which will be reused for training, development and evaluation data will consist of **any kind of speech found in the Internet** (YouTube videos),

GTTS Tecnologías Software

Future work

Actually, current work: **KALAKA-3**

- Support for the **Albayzin 2012 Language Recognition Evaluation**: June to October 2012, results presented at **IberSpeech 2012**, to be held in Madrid (Spain) in November 2012.
- Besides including all the materials of KALAKA-2, which will be reused for training, development and evaluation data will consist of **any kind of speech found in the Internet** (YouTube videos), **4 new target languages** will be added: French, German, Greek and Italian,

Future work

Actually, current work: **KALAKA-3**

- Support for the **Albayzin 2012 Language Recognition Evaluation**: June to October 2012, results presented at **IberSpeech 2012**, to be held in Madrid (Spain) in November 2012.
- Besides including all the materials of KALAKA-2, which will be reused for training, development and evaluation data will consist of **any kind of speech found in the Internet** (YouTube videos), **4 new target languages** will be added: French, German, Greek and Italian, and **many other OOS languages** will be also recorded.

Future work

Actually, current work: **KALAKA-3**

- Support for the **Albayzin 2012 Language Recognition Evaluation**: June to October 2012, results presented at **IberSpeech 2012**, to be held in Madrid (Spain) in November 2012.
- Besides including all the materials of KALAKA-2, which will be reused for training, development and evaluation data will consist of **any kind of speech found in the Internet** (YouTube videos), **4 new target languages** will be added: French, German, Greek and Italian, and **many other OOS languages** will be also recorded.

More info at http://iberspeech2012.ii.uam.es/ (under Albayzin Evaluations)

GTTS Tecnologías Software

Future work

Actually, current work: **KALAKA-3**

- Support for the **Albayzin 2012 Language Recognition Evaluation**: June to October 2012, results presented at **IberSpeech 2012**, to be held in Madrid (Spain) in November 2012.
- Besides including all the materials of KALAKA-2, which will be reused for training, development and evaluation data will consist of **any kind of speech found in the Internet** (YouTube videos), **4 new target languages** will be added: French, German, Greek and Italian, and **many other OOS languages** will be also recorded.

More info at http://iberspeech2012.ii.uam.es/ (under Albayzin Evaluations)

You are all invited to participate !!!

GTTS Tecnologías Software