KALAKA-3:
a database for the recognition of spoken European languages on YouTube audios

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

Grupo de Trabajo en Tecnologías Software (GTTS, http://gtts.ehu.es)
University of the Basque Country UPV/EHU, Spain
e-mail: luisjavier.rodriguez@ehu.es

LREC 2014
Reykjavik (Iceland)
May 28-30 2014

Outline

- Spoken language recognition
- Albayzin LRE 2012
- KALAKA-3: main features
- KALAKA-3: design and collection procedure
- KALAKA-3: evaluation
- Conclusions and future work

Spoken Language Recognition

- **Is utterance X spoken in French?**
- **Give me a score** (the higher the score, the higher the likelihood that X is spoken in French)
- To make a decision, **apply a threshold** to the given score

SLR Evaluation

- **Performance**: decisions compared to ground-truth for a set of speech files and target languages
- **Types of tests**:
  - closed-set (known set of target languages)
  - open-set (any language could be spoken)
- **Difficulty**:
  - background and/or channel conditions
  - dialect variability
  - short utterances

International SLR Benchmarks

- **NIST LRE**:1996, 2003, 2005, 2007, 2009 and 2011
  - Focused on telephone speech for large-scale filtering in security applications, dealing with certain languages of interest (for strategic reasons)

- **Albayzin LRE**: 2008, 2010 and 2012
  - Initially dealing only with languages spoken in Spain, then extended to other European languages
  - 2008 LRE run on clean Broadcast News (BN) speech
  - 2010 LRE run on BN speech with noisy segments
  - 2012 LRE run on unrestricted speech found in Internet (YouTube audios)

Albayzin 2012 LRE

- Designed to address the conditions producing **variability or difficulty** in previous evaluations
  - Unconstrained speech (background, channel, dialect, amount of speech available, etc.)
  - Low-resource scenario (few data available)
- **Target application**: indexing the spoken language in multimedia contents
- Task defined this way was of practical interest and challenging enough to foster research

KALAKA-3: main features

- Created to support the Albayzin 2012 LRE
- Recycles **BN speech** from previous evaluations (for training: 6 target languages)
- Includes **unconstrained speech** signals from YouTube videos (for tuning and testing)
- **Tasks**:
  - **Plenty-of-Training**: 6 target languages
  - **Empty-Training**: 4 target languages
- Open-set tests: 11 Out-Of-Set (OOS) languages

KALAKA-3: main features

- Three datasets: Train, Dev and Eval
- **Train**: 108 hours, 18 hours per target language (80% clean, 20% noisy)
- **Dev/Eval**: same size (**+2000 YouTube audios**), target languages balanced, different distribution of OOS languages
- KALAKA-3: **~200 hours**, currently distributed as a set of tarballs (for downloading), after direct request to authors

KALAKA-3: design

- **Goal**:
  - 300 YouTube videos per target language (150 Dev + 150 Eval)
  - 100 YouTube videos per OOS language
- **Dev/Eval** datasets as **independent** as possible, to avoid a biased benchmark
- Duration: **30-120 seconds**, including **at least 5 seconds of speech**
- Audios with **telephone speech discarded**

KALAKA-3: collecting data

- **(1) Lists (spreadsheets) of candidate YouTube videos** automatically created for each language

  - **list of language-specific keywords**:
    - ✓ 2000 words (canonical forms) with 6 or more characters randomly chosen from the aspell dictionary
    - ✓ words in the aspell dictionary of other language excluded
    - ✓ only **1000 keywords** retained per language
  - **6 YouTube categories most likely to contain speech**: Education, News, Entertainment, Howto, Nonprofit, Technology
  - For each (language, category), list of videos built by **filtering per category and duration** and **searching for keywords** in metadata, **using YouTube API v2.0**

KALAKA-3: collecting data

- **(2) Videos ranked in spreadsheets according to:**
  - **Creative Commons (CC) license** (not many)
  - **Geographical location** (geographical metadata not always available):
    - priority given to videos located within a certain distance from a major city speaking the language of interest
    - a small list of major cities defined for each language
    - distance depending on the size of the country (typically, R = 200 km)

KALAKA-3: collecting data

- **(3) Validation**
  - Each (language, category) spreadsheet scrolled through and annotated with validation marks
  - **Videos listened to and subjectively judged by 5 human auditors (2 months)**
  - **Videos validated in order**, until the desired amount (55 for target languages, 17 for OOS languages) is attained
  - **A video is validated if and only if**:
    - ✓ contains **+5 seconds of speech**
    - ✓ contains speech in a **single language** (for OOS languages, several languages may appear, but not target languages)
    - ✓ **background/channel conditions are admissible**

KALAKA-3: collecting data

- **(4) Fetching and converting YouTube audios**
  - Videos fetched using **youtube-dl**
  - Audio layer extracted using **ffmpeg**
  - Audio converted to single-channel 16-kHz 16-bit PCM encoded WAV files using **SoX**
  - **Filenames anonymized**
  - The database provides no information about the original videos (only the spoken language is given in the ground-truth files)

KALAKA-3: YouTube video collection

- 4168 audios validated out of 21860 audited
- **Dev**: 2059 (News, Education, Howto)
- **Eval**: 2019 (Entertainment, Nonprofit, Technology)
- At least 150 videos per target language
- Different OOS distribution

		Devel	Eval
Target languages (Plenty-of-Training)	**Basque**	154	150
	Catalan	149	158
	English	150	156
	Galician	151	160
	Portuguese	160	163
	Spanish	153	154
Target languages (Empty-Training)	**French**	150	155
	German	146	151
	Greek	155	165
	Italian	158	160
OOS languages	**Bulgarian**	0	98
	Croatian	90	0
	Czech	102	0
	Finnish	0	89
	Hungarian	51	51
	Polish	102	0
	Romanian	98	0
	Russian	45	54
	Serbian	0	91
	Slovak	0	102
	Ukrainian	45	52

KALAKA-3: evaluation

- **New metric**: F_{act} (actual relative confusion), ranging between 0 (perfect system) and 1 (non-informative system)
- **Task reformulated**: given an audio X and N target languages, systems must provide N+1 scores (for target and OOS languages)
- PO performance only slightly worse than PC: low confusion between target and OOS languages (**design flaw**)
- EC/EO performance much worse than PC/PO (late systems 1 and 6 used dev data for training): **lack of training data is a challenging condition !!!**

Albayzin 2012 LRE: Summary of results

Systems	PC	PO	EC	EO
1	**0.071**	**0.085**	–	–
2	0.078	0.120	0.498	0.516
3	0.113	0.114	0.711	0.796
4	0.121	0.160	0.626	0.676
5	0.122	–	–	–
6	0.141	0.184	–	–
7 (late)	0.407	0.216	–	–
1 (late)	–	–	**0.216**	–
6 (late)	–	–	0.310	**0.372**

KALAKA-3: evaluation

- **Acoustic SLR systems** with competitive performance on other tasks (NIST LRE): MFCC/SDC-iVector and PLLR-iVector

- **Basic Voice Activity Detection (VAD)** based on PLLRs (could be failing due to background music or conversations)

- **1/2 development data used for training** in the EC/EO tracks (note that dev data were not intended for training)

- Performance comparatively good in EC/EO, but not in PC/PO (VAD errors, lack of phonotactic systems)

Results for two acoustic SLR systems and the fusion of them

Systems	PC	PO	EC	EO
iVector-MFCC	0.139	0.254	0.238	0.342
iVector-PLLR	0.191	0.294	0.217	0.341
Fusion	0.098	0.128	0.131	0.221

Conclusions and future work

- KALAKA-3 provides challenging tasks for the development of SLR technology
- As far as we know, this is the first SLR benchmark dealing with unconstrained speech found in Internet (YouTube audios)
- Already used by several research groups
- We plan to license the Dev and Eval datasets through LDC