

Audio Surveillance Tapping Transcription Quality Estimation Reliability

Clara Borrelli, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, Stefano Tubaro
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
Email: {clara.borrelli, paolo.bestagini, fabio.antonacci, augusto.sarti, stefano.tubaro}@polimi.it

Abstract—The abstract goes here.

I. INTRODUCTION

Thanks to the recent advances in technology, audio communication systems are becoming more and more pervasive and diverse. In fact, nowadays a conversation can happen not only face-to-face but also on a variety of channels, like phone calls, VoIP or voice messages on instant messaging platforms.

At the same time the ability of monitoring environmental, digital or phone communications has become an urgent necessity for national security issues. These technologies can effectively assist law enforcement agencies in foiling terrorist attacks or revealing harmful intents. The aforementioned heterogeneity of the communication channels has produced in this sense both advantages and drawbacks.

On one side, information gathering has become easier and ubiquitous. Regarding environmental monitoring, audio surveillance devices are getting cheaper and easier to deploy. Subsequently, the number of control spots can be further increased, ensuring higher coverage, for example, in large public spaces. Concerning digital or phone wire taps, *why is easier to collect data? because we have so many communication channels?*

On the other side, a blind and massive data collection can result in huge databases whose manual inspection might be infeasible. Tools like automatic transcription agents would be of great help for both investigative and surveillance purposes. However, given the diversity of context in which data is collected, the audio excerpts can be affected by several types of noise which can degrade the quality of the records and compromise the intelligibility of a possible relevant conversation.

For these reasons it is needed to develop automatic and intelligent methods that allow to speed up the analysis of these huge corpora but at the same time take in account the variety of the involved devices and environments. Depending on the characteristics of the collected audio excerpts, these systems should be able to extract relevant information while guaranteeing control to the human user.

In this work we try to meet these needs by proposing a framework able to evaluate the quality of noisy speech recordings. This evaluation is based on the estimation of

the likelihood of obtaining reliable transcript using a generic automatic speech-to-text engine.

Our method can be employed in two different applicative contexts. The first one is for *not sure what is the meaning of the bullet point. We want to propose an index to be exploited for tuning automatic transcription agents? or we want to compute the intelligibility index on large databases to evaluate the quality of a specific dataset?*. The second one is, as already mentioned, to provide a useful tool in managing data acquired by audio surveillance systems. The model will provide a quality feedback to investigators, that will be able to discriminate reliable from non reliable automatically extracted transcriptions.

- Audio communication systems are becoming more pervasive and diverse (e.g., phone calls, voip, voice messages, etc.).
- For national security reason (e.g., to avoid terroristic attacks, uncover malicious things, etc.), the ability of monitoring communications through phone tapping or environmental monitoring in several situations is an urgent necessity.
- Fortunately, audio surveillance devices are increasingly cheaper and easier to deploy.
- However, a blind and massive data collection might result in huge databases whose analysis might become unfeasible.
- It is therefore needed to develop methods that automatically enable to speed the analysis of these huge corpora of collected audio excerpts.
- In this paper we propose a framework to evaluate the quality of noisy speech recordings estimating the likelihood of obtaining a reliable transcript.
- Two applications: i) to process large databases with automatic speech-to-text transcribers; ii) to provide a quality feedback to any investigator deploying an audio surveillance system.
- The main rationale is to use ...
- Results obtained on a lots of data show that ...

[1]

II. PROBLEM STATEMENT AND BACKGROUND

In this section bla bla ...

A. Problem formulation

Given an audio excerpt $x(t)$ composed by ... containing some speech ... we want to estimate how much it is useful for an analyst. This is done by estimating a score indicating how likely is it possible to correctly understand the speech.

B. Background

Ci sono tanti metodi in tante aree che si occupano di stima dell'intelligibilit . Citiamono un po' e facciamo capire che sono tipicamente cuciti su scenari applicativi per noi non rilevanti (e.g., acustica di stanza, trasmissioni telefoniche, etc.).

III. METHOD

Cappello introduttivo in cui si spiega che il framework   composto da tre elementi

- La scelta di una misura di qualita' oggettiva (Jaccard, Coseno su vettori di parole)
- L'estrazione di descrittori (MFCC, etc.)
- Classificazione / regressione

A. Objective quality measure

B. Feature extraction

C. Decision taking

IV. RESULTS

In this section bla bla ...

A. Dataset

Da dove siamo partiti, tipi di rumore, SNR, trascrittore.

B. Setup

Tipi di classificatore / regressore usati. Divisione train, validation e test (o forse in dataset).

C. Numerical analysis

risultati di classificazione / regressione con metodi diversi, e feature diverse.

V. CONCLUSIONS

It worked!

ACKNOWLEDGMENT

This material is based on research sponsored by DARPA and Air Force Research Laboratory (AFRL) under agreement number FA8750-16-2-0173. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA and Air Force Research Laboratory (AFRL) or the U.S. Government.

REFERENCES

- [1] A. Piva, "An overview on image forensics," *ISRN Signal Processing*, 2013.