SPEECH TECHNOLOGY RESEARCH AT LPTV

Prof. Néstor Becerra Yoma, Ph.D.

Laboratorio de procesamiento y transmisión de voz
(Speech Processing and Transmission Lab)
Universidad de Chile
nbecerra@ing.uchile.cl

http://www.lptv.cl
http://www.cmrsp.cl

DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
UNIVERSIDAD DE CHILE

Speech Processing and Transmission Laboratory

Speech Processing and Transmission Laboratory

- LPTV (Laboratorio de Procesamiento y Transmisión de Voz) was started in 2000.

- Carry out R&D on robust speech recognition/speaker verification, CAPT (computer aided pronunciation training), CALL (computer aided language learning), dialogue systems and voice over IP and, more recently, voice-based human robot interaction.

Speech Processing and Transmission Laboratory

- At that point, there had been some efforts in Latin America to carry out R&D in speech technology.

- However, this activity had hardly reached the international community.

Speech Processing and Transmission Laboratory

Old times….2000 or 2001

DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
UNIVERSIDAD DE CHILE

Speech Processing and Transmission Laboratory

Uncertainty in noise cancelling

Uncertainty in noise cancelling was firstly proposed to weight the information provided by frames according to their reliability in DTW and HMM algorithms in 1994-1998.

N. Becerra Yoma, F. R. McInnes and M. A. Jack, "Improving Performance of Spectral Subtraction in Speech Recognition Using a Model for Additive Noise," IEEE Transactions on Speech and Audio Processing, vol. 6, no. 6, pp. 579-582, 1998

N. Becerra Yoma, F. R. McInnes and M. A. Jack, "Improving Performance of Spectral Subtraction in Speech Recognition Using a Model for Additive Noise," IEEE Transactions on Speech and Audio Processing, vol. 6, no. 6, pp. 579-582, 1998

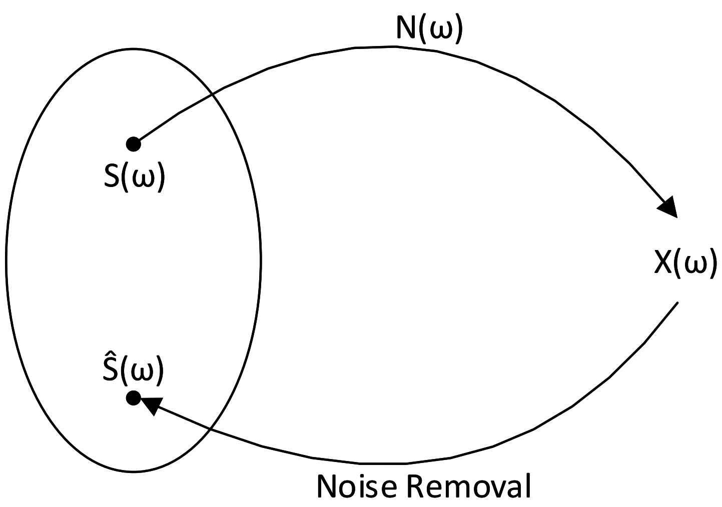
N. Becerra Yoma, F. R. McInnes and M. A. Jack, "Weighted Matching Algorithms and Reliability in Noise Canceling by Spectral Subtraction," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 97, Munich, Germany, 1997

Uncertainty in noise cancelling

Model

x(t)=s(t)+n(t)

s(t), clean signal
n(t), additive noise
x(t), noisy signal



Uncertainty in noise cancelling

Model:

$$\overline{x_m^2} = \overline{s_m^2} + \overline{n_m^2} + 2 \cdot \sqrt{c_m} \sqrt{\overline{s_m^2}} \cdot \sqrt{\overline{n_m^2}} \cdot \cos(\phi)$$

$$E\left[\log(\overline{s_m^2}(\phi))\right] = \int_{-\pi}^{\pi} \log(\overline{s_m^2}(\phi)) \cdot f_\phi(\phi) \cdot d(\phi) \cong \log(E[B_m])$$

Where

$$B_m = \overline{x_m^2} - \overline{n_m^2}$$

$$E[B_m] = \overline{x_m^2} - E\left[\overline{n_m^2}\right]$$

Uncertainty in noise cancelling

Consequently, the uncertainty variance can be defined as:

$$Var\left[\log(\overline{s_m^2}(\phi))\right] = E\left[\log^2(\overline{s_m^2}(\phi))\right] - E^2\left[\log(\overline{s_m^2}(\phi))\right]$$

$$Var\left[\log\left(\overline{s_m^2}(\phi)\right)\right] \cong \frac{2 \cdot c_m \cdot E\left[\overline{n_m^2}\right]}{\overline{x_m^2} - E\left[\overline{n_m^2}\right]}$$

But, how could it be used?

Stochastic Weighted Viterbi algorithm

If the features extracted from the speech signal are random variables, what would it happen with the ordinary HMM observation probability?

In most HMM systems the output probability is modeled with a mixture of Gaussians with diagonal covariance matrices.

We proposed to replace the expected value of the output probability if the features are considered random variables with Gaussians distributions.

N. Becerra Yoma and M. Villar, "Speaker verification in noise using a stochastic version of the weighted Viterbi algorithm," *IEEE Transactions on Speech and Audio Processing,* vol. 10, no. 3, pp. 158-166, 2002.

Stochastic Weighted Viterbi algorithm

This is the first time that this idea was published, but some authors do not like to acknowledge that.

If the HMM observation probability is defined as:

$$b_s(O_t) = \sum_{g=1}^{G} p_g \cdot \prod_{n=1}^{N} (2\cdot\pi)^{-0.5} \cdot (Var_{s,g,n})^{-0.5} \cdot e^{-\frac{1}{2}\cdot\frac{(O_{t,n}-E_{s,g,n})^2}{Var_{s,g,n}}}$$

Stochastic Weighted Viterbi algorithm

Then, the expected value of the HMM observation probability is given by:

$$E[b_s(O_t)] = \sum_{g=1}^{G} p_g \cdot \prod_{n=1}^{N} \frac{1}{\sqrt{2 \cdot \pi \cdot Vtot_{s,g,n,t}}} \cdot e^{-\frac{1}{2} \cdot \frac{(E[O_{t,n}] - E_{s,g,n})^2}{Vtot_{s,g,n,t}}}$$

where

$$Vtot_{s,g,n,t} = Var_{s,g,n} + Var(O_{t,n})$$

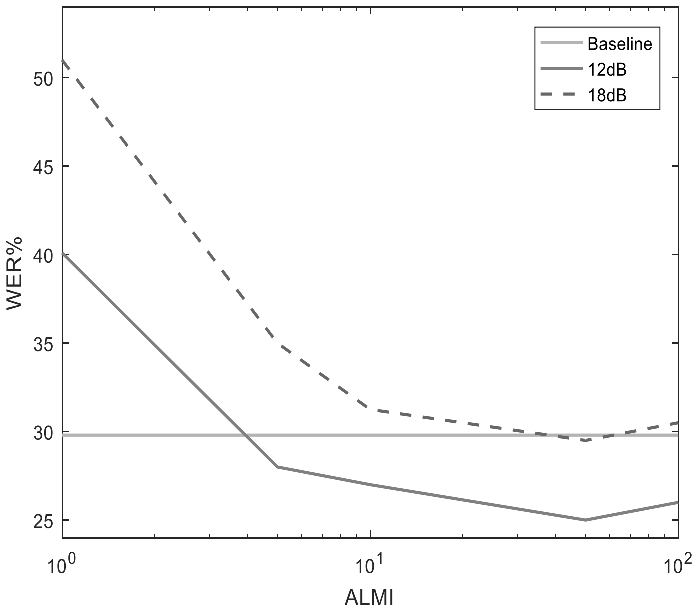
$Var(O_{t,n})$ is the uncertainty variance.

$Var_{s,g,n}$ is the HMM variance

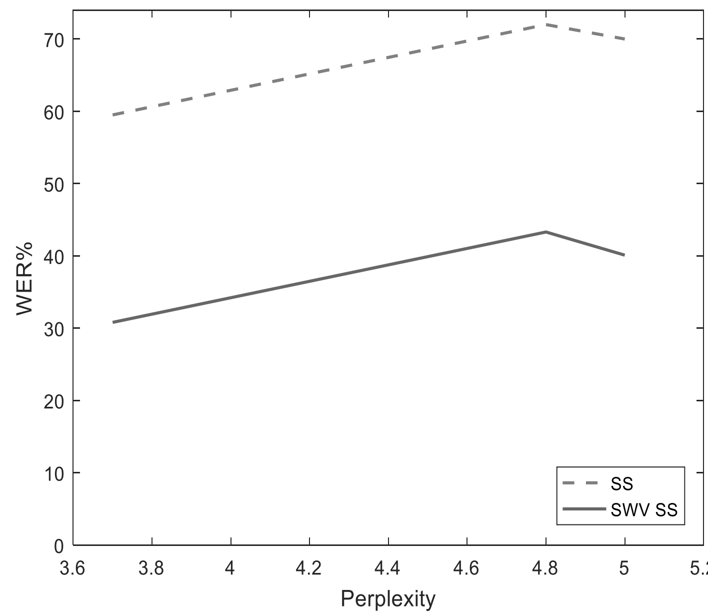
DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
UNIVERSIDAD DE CHILE

Speech Processing and Transmission Laboratory

Stochastic Weighted Viterbi algorithm



WER vs. ALMI with a trigram LM



WER vs LM perplexity with car noise, SNR=12dB

Weighted Viterbi algorithm with DNN

What about deep learning?

We proposed a similar scheme for HMM-DNN based decoding:

$$\widehat{W} = \arg\max_{W} \{UW \cdot \log[p(X|W)] + \lambda \cdot \log[p(W)]\}$$

X denotes the sequence of acoustic observations x_t, and $p(X|W)$ is the acoustic model probability that depends on the pseudo log-likelihood delivered by the DNN, $\log[p(x_t|q)]$.

$p(W)$ is the language model probability of word string W and λ is the scaling factor.

.

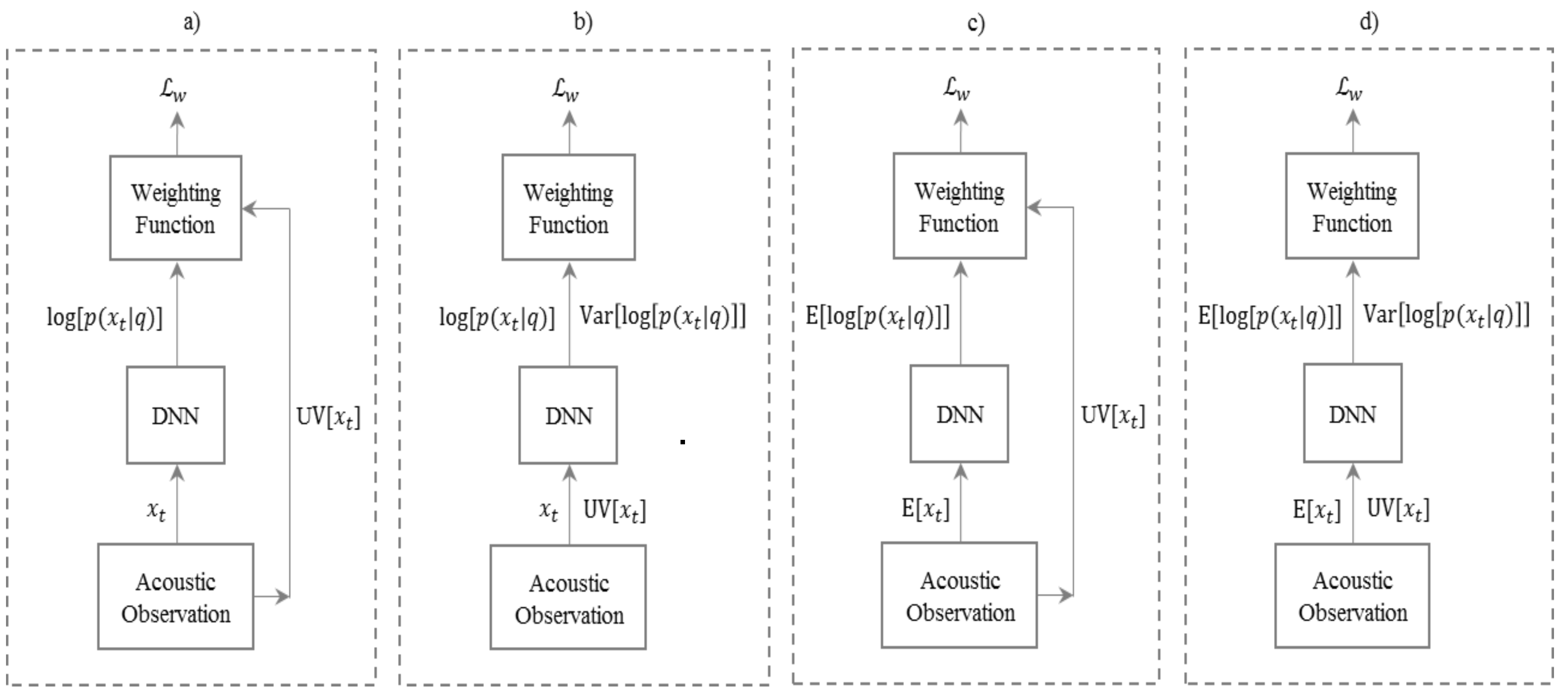
Jose Novoa, Jose Fredes, Nestor Becerra Yoma. "Uncertainty weighting and propagation in DNN-HMM-based speech recognition". Submitted to Computer, Speech and Language, September 2016.

DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
UNIVERSIDAD DE CHILE

Speech Processing and Transmission Laboratory

Weighted Viterbi algorithm with DNN

Proposed uncertainty weighting scheme with and without uncertainty propagation. The weighted pseudo log-likelihoods, $\mathcal{L}_w$, corresponds to: $UW[x_t] \cdot \log[p(x_t|\mathrm{q})]$ in a) and b); and $UW[x_t] \cdot \mathrm{E}\{\log[p(x_t|q)]\}$ in c) and d).



fcfm DEPARTAMENTO DE INGENIERIA ELECTRICA FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS UNIVERSIDAD DE CHILE

Speech Processing and Transmission Laboratory

State duration modeling

Original HMM state duration distribution is geometric. We also proposed state duration modelling for HMM to model better speech, Internet packet-lost and volcano events.

In Arabic language duration is used to discriminate some phonemes.

Néstor Becerra Yoma and Jorge Silva. "MAP speaker adaptation of state duration distributions for speech recognition." IEEE Transactions on Speech and Audio Processing Vol. 10, Nº7, pp.443-450, 2002.

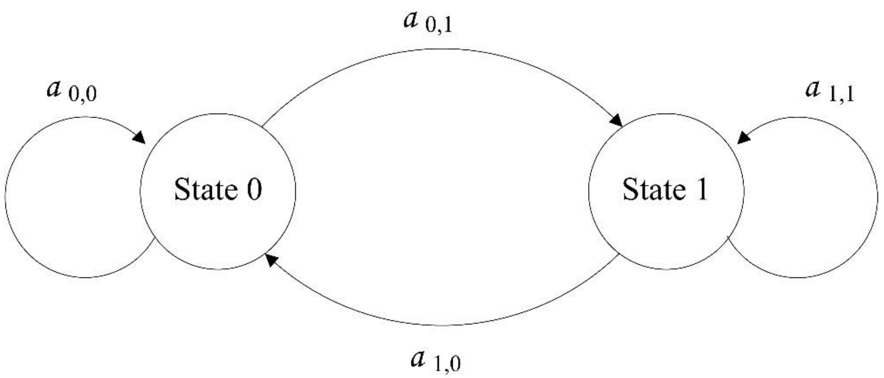
Néstor Becerra Yoma and Tarciano F. Pegoraro. "Robust speaker verification using state duration modeling." Speech Communications (Elsevier), Vol. 38, pp.77-88, 2002.

Néstor Becerra Yoma, Fergus McInnes, Mervyn Jack, Sandra Stump, Lee Luan Ling. "On including temporal constraints in the Viterbi algorithm for speech recognition in noise." IEEE Transactions on Speech and Audio Processing, Vol. 9, No.2, pp. 179-182, 2001.

VoIP

10 or even 15 years ago, VoIP was a new paradigm.

We developed a method to assess que subjective quality of the speech transmitted on IP networks:

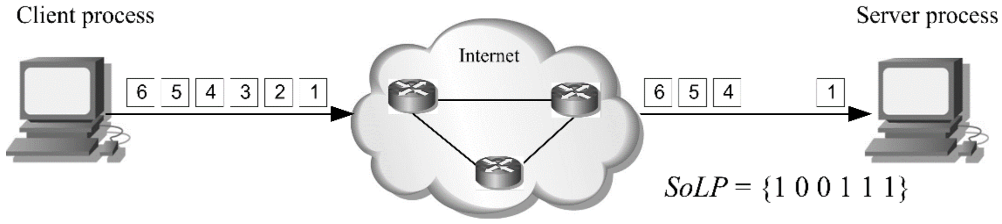


We included state duration constraints to model the packet-loss process in IP networks more accurately.

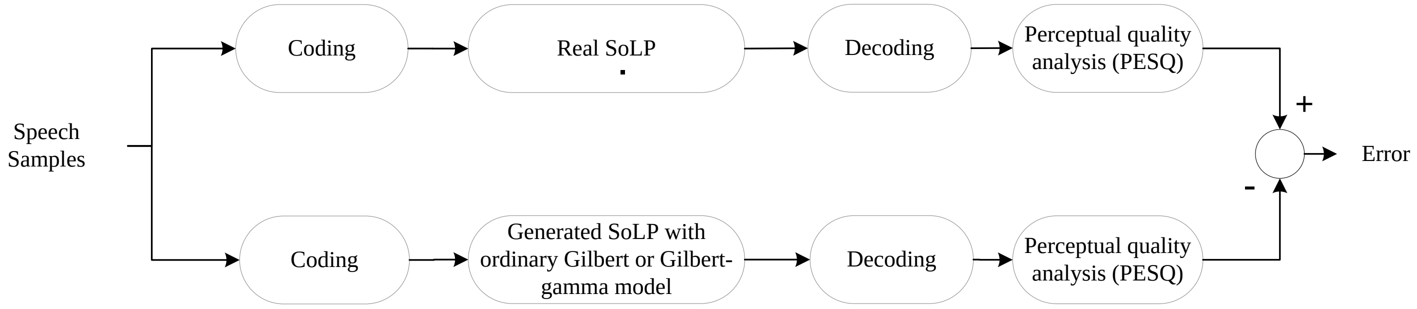
Nestor Becerra Yoma, Carlo Busso, Ismael Soto. "Packet-loss modeling in IP networks with state duration constraints." IEE Proceedings on Communications, Vol. 152 (1), pp.1-5, 2005.

VoIP

Procedure to generate Sequence of lost packets

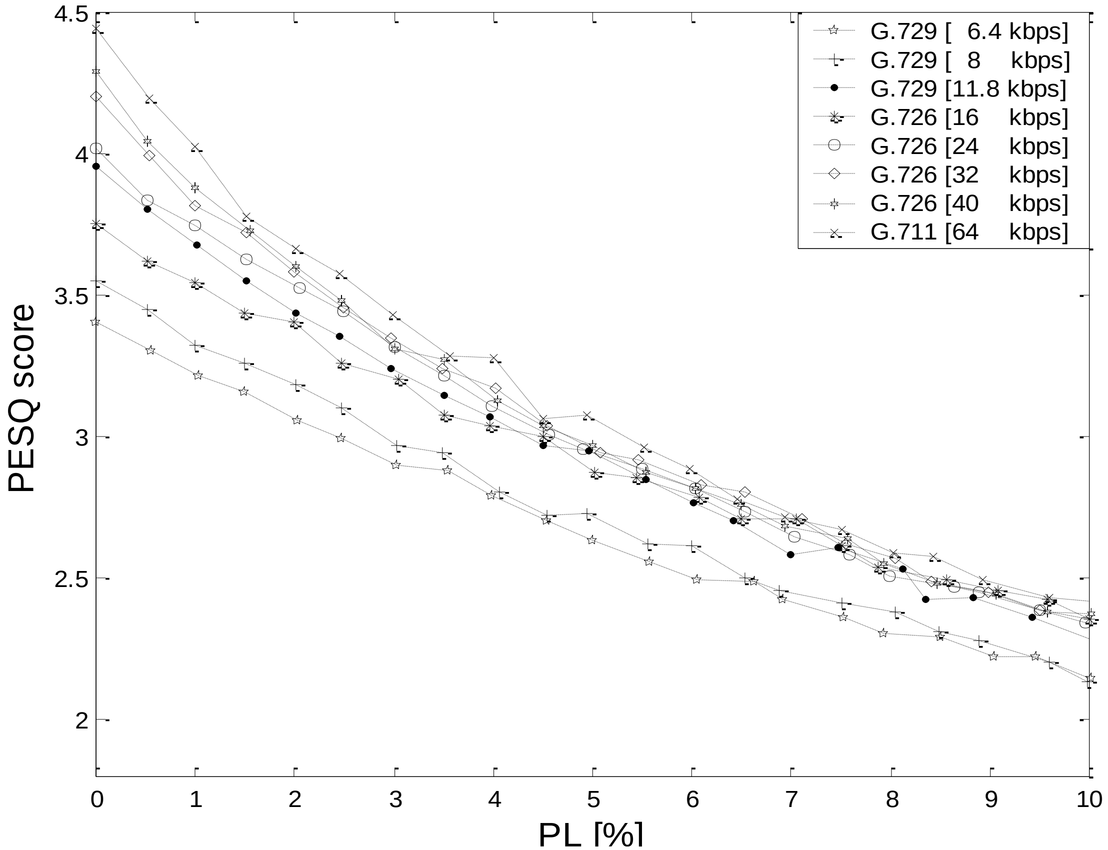


Evaluation of the packet-loss model



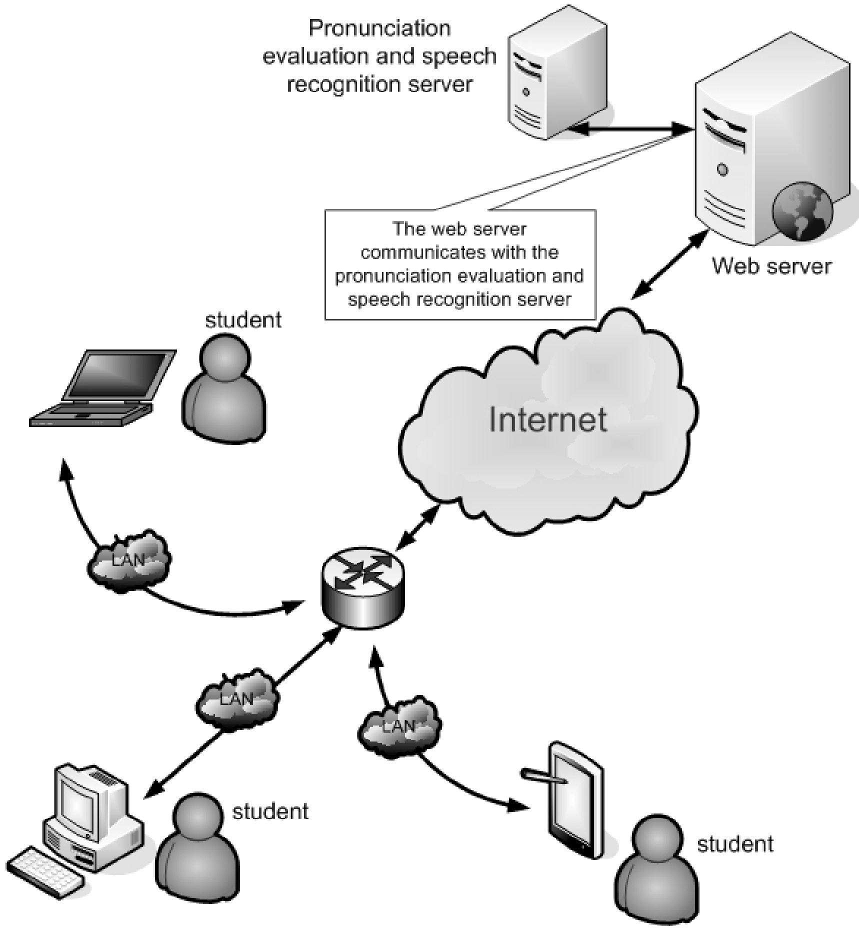
VoIP

Procedure to generate Sequence of lost packets



CAPT in the cloud … but in 2007!

Deployment of
a CAPT application
with a central server



CAPT in the cloud … but in 2007!

In the framework of the remote processing in the previous slide we proposed:

-A method to automatically generate the competitive lexicon, required by an ASR engine to compare the pronunciation of a target word with its correct and wrong phonetic realizations.

-A Bayes-based multi-classifier fusion approach to map ASR objective confidence scores to subjective evaluations in pronunciation assessment is presented.

Carlos Molina, Néstor Becerra Yoma, Jorge Wuth, Hiram Vivanco. "ASR based pronunciation evaluation with automatically generated competing vocabulary and classifier fusion." Speech Communciations (Elsevier), 51(2009), pag. 485-498, 2009.

DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
UNIVERSIDAD DE CHILE

Speech Processing and Transmission Laboratory

CAPT in the cloud … but in 2007!

Distance between two words with non-linear alignment of models

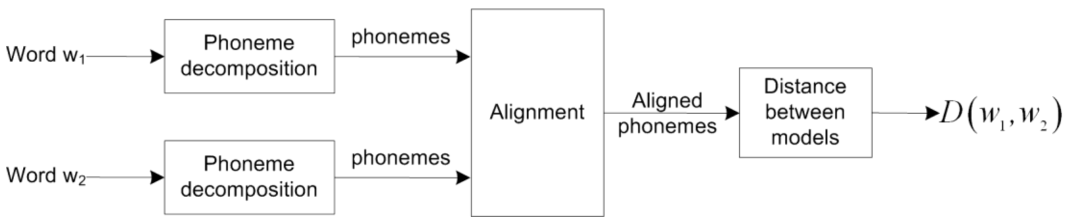
$$D\left(w_x, w_y\right) = \frac{1}{K}\sum_{k=1}^{K} d\left(\lambda_x^{m_x(k)}, \lambda_y^{m_y(k)}\right)$$

Spanish	English	Spanish	English
a$_s$	Ah	m$_s$	**M**
b$_s$	B	n$_s$	**N**
ch$_s$	Ch	o$_s$	**Aa**
d$_s$	D, Dh	p$_s$	**P**
e$_s$	Eh	r$_s$	**R**
f$_s$	F	s$_s$	**S**
g$_s$	G	t$_s$	**T**
i$_s$	Ih	u$_s$	**Uh**
j$_s$	Hh	w$_s$	**W**
k$_s$	K	x$_s$	**KS**
l$_s$	**L**	**y$_s$**	**Y**

DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS UNIVERSIDAD DE CHILE

Speech Processing and Transmission Laboratory

CAPT in the cloud … but in 2007!



w_t	s	ay	ah	n	t	ah	s	t	English phoneme decomposition
$w_{t\text{-}SP}$	s_s	i_s	e_s	n_s	t_s	i_s	s_s	t_s	Spanish phoneme decomposition using spanish units
									Phoneme replacement
$w_{t\text{-}EP}$	s	ih	eh	n	t	ih	s	t	Spanish phoneme decomposition using english units

CAPT in the cloud … but in 2007!

Provided a lexicon automatically generated from the target pronunciation, the ASR engine can deliver the following metrics:

Word density confidence measure

$$WDCM_t = \frac{\sum_{r \in E(w_t, H)} Q(h_r)}{\sum_{l=1}^{N} Q(h_l)}$$

Recognition flag

$$REC_t = \begin{cases} 1 & \text{if } w_t \subset h_1 \\ 0 & \text{if } w_t \not\subset h_1 \end{cases}$$

Maximum hypothesis log-likelihood

$$ML_t = log\left[\max_r [Q(h_r)] \mid r \in E(w_t, H) \right]$$

Difference between maximum and minimum state duration

$$MmSD_t = MaxSD(t) - MinSD(t)$$

Position in the N-best

$$POS_t = arg\max_r \left\{ [Q(h_r)] \mid r \in E(w_t, H) \right\}$$

DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
UNIVERSIDAD DE CHILE

Speech Processing and Transmission Laboratory

CAPT in the cloud … but in 2007!

The subjective score given a metric can be estimated as:

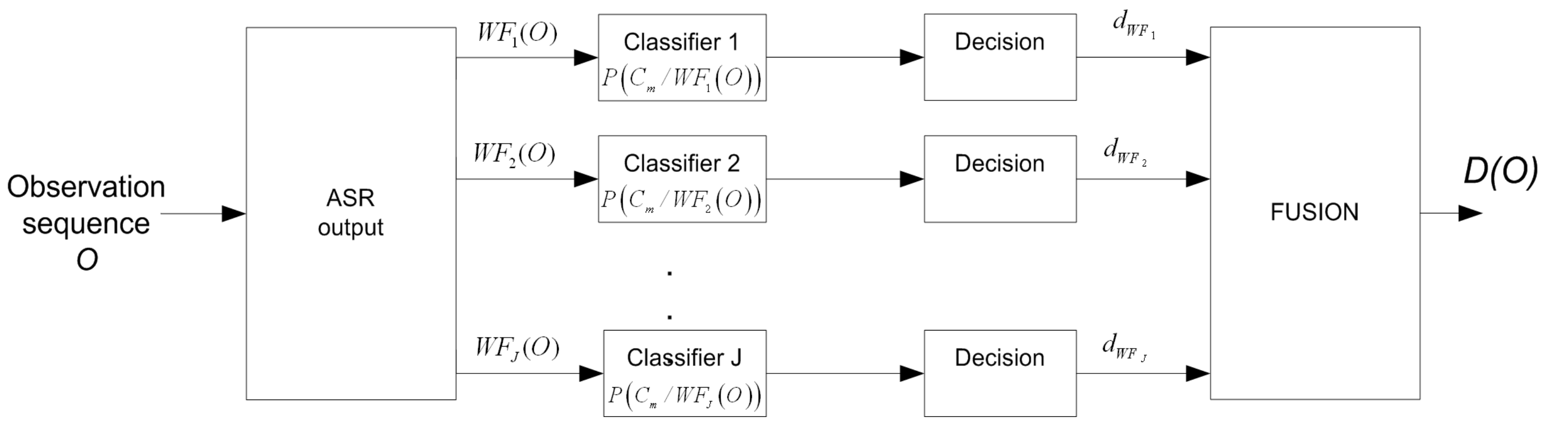
$$d_{WF_j}(O) = \arg\max_{C_m} P\left(C_m / WF_j(O)\right) = \arg\max_{C_m} \left\{ \frac{P\left(WF_j(O) / C_m\right) \cdot P(C_m)}{P\left(WF_j(O)\right)} \right\}$$

Multi-classifier fusion at feature level:

$$D(O) = \arg\max_{C_m} P\left(C_m / \overrightarrow{WF}(O)\right) = \arg\max_{C_m} \left\{ \frac{P\left(\overrightarrow{WF}(O) / C_m\right) \cdot P(C_m)}{P\left(\overrightarrow{WF}(O)\right)} \right\}$$

CAPT in the cloud … but in 2007!

Multi-classifier fusion at abstract level:



CAPT in the cloud … but in 2007!

DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS UNIVERSIDAD DE CHILE

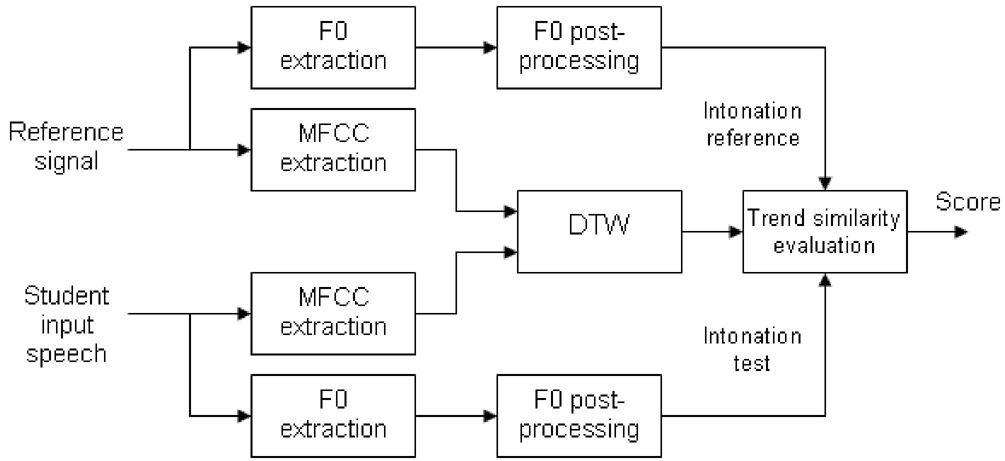
Speech Processing and Transmission Laboratory

What about intonation?

In several languages there is not only one correct intonation pattern.

However, this result has also been used in Chinese Mandarin.

We also used intonation modeling for emotion detection.



J.P. Arias, N. Becerra Yoma, H. Vivanco."Automatic intonation assessment for computer aided language learning." Speech Communications (Elsevier), Vol. 52, Issue 3, March 2010, pages 254-267.

J.P. Arias, C. Busso, N. Becerra Yoma. "Shape-based modeling of the fundamental frequency contour for emotion detection in speech." CSL(Elsevier), Volume 28, Issue 1, January 2014, Pages 278–294.

Confidence-based classifier fusion

We proposed a method based on the maximization of the Bayes-based confidence for multi-classifier fusion:

$$BBCM(WF_i) = Pr(w_i \text{ is ok}|WF_i) = \frac{Pr(WF_i \mid w_i \text{ is OK}) \cdot Pr(w_i \text{ is OK})}{Pr(WF_i)}$$

BBCM: Bayes-based confidence measure.

Besides speech, this paper has also been cited in fields such as fault detection, natural language processing, image recognition, bio-engineering and multi-media big data retrieval.

Fernando Huenupan, Néstor Becerra Yoma, Claudio Garretón, Carlos Molina. "Confidence based multiple classifier fusion in speaker verification." Pattern Recognition Letters, Vol 29/7 pp 957-966, 2008.

Locally normalized filter bank (LNFB)

We proposed a set of features that is able to remove coarse variations in the spectral shape on a frame-by-frame basis.

The Seneff´s GSD (generalized synchrony detector) is given by:

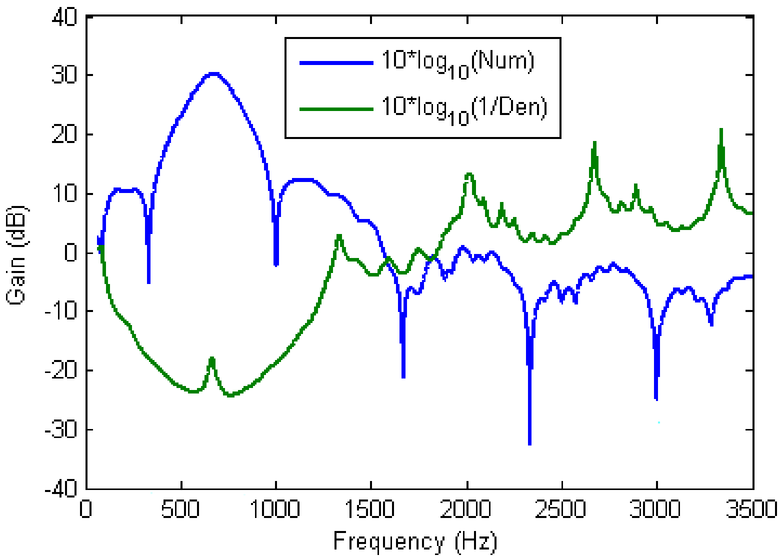
$$GSD_i(y) = A_s \tan^{-1}\left[\frac{1}{A_s}\left(\frac{\langle|y[n]+y[n-n_i]|\rangle-\delta}{\langle|y[n]-\beta^{n_i}y[n-n_i]|\rangle}\right)\right]$$

GSD is a component of auditory models that had been widely explored with limited success.

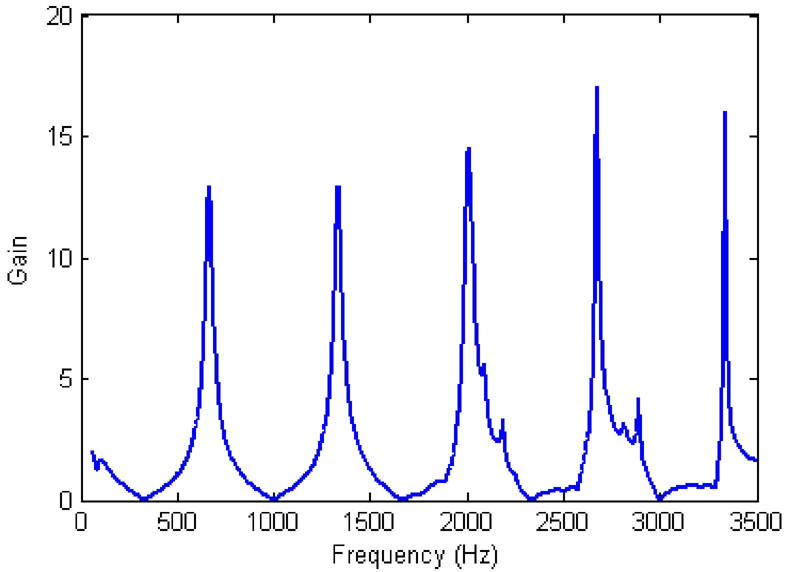
Victor Poblete, Felipe Espic, Simon King, Richard M. Stern, Fernando Huenupán, Josué Fredes, Néstor Becerra Yoma. "A perceptually-motivated low-complexity instantaneous linear channel normalization technique applied to speaker verification." CSL (Elsevier) , Vol.31, N°1 May 2015, Pages 1–27.

Locally normalized filter bank (LNFB)

We interpreted the GSD model as a filter and estimated its transfer function:



Close up of the log magnitude frequency response of the numerator and denominator of the GSD tuned at 692(Hz). The numerator $\mathbf{10 * \log_{10}(\langle|y[n] + y[n-n_i]|\rangle - \delta)}$ is in blue and the denominator $\mathbf{10 * \log_{10}(1/\langle|y[n] - \beta^{n_i} y[n-n_i]|\rangle)}$ in green.



Frequency response of the GSD tuned at 692(Hz).

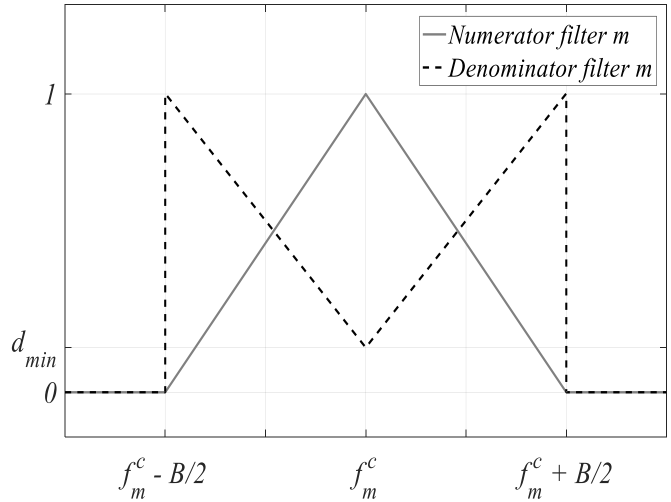
DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS UNIVERSIDAD DE CHILE

Speech Processing and Transmission Laboratory

Locally normalized filter bank (LNFB)

To avoid the GSD spurious periodic response, we proposed the locally normalized filter bank (LNFB) based on the following set of filters:

$$LNFB_m = \log(LN_m) = \log(LNNum_m / LNDen_m)$$



LNFB in Aurora-4

Training	Aurora Group	Mel FB + ΔMelFB	LNFB + ΔLNFB2
Clean (As in Group A)	A	2.39	2.65
	B	19.70	19.26
	C	21.69	14.10
	D	39.55	34.02
	Average	27.12	24.03
Multi-Noise (As in Group B)	A	2.56	3.19
	B	6.11	6.94
	C	16.57	11.84
	D	25.06	21.70
	Average	14.73	13.35
Multi-condition (As in Group D)	A	3.42	3.62
	B	6.35	7.19
	C	7.12	7.38
	D	16.68	17.06
	Average	10.62	11.18
All-Condition Average		17.5	16.2

Josue Fredes et al. "Locally-Normalized Filter Banks Applied to Deep Neural Network-based Robust Speech Recognition." Accepted for publication in IEEE Signal Processing Letters, January 2017.

fcfm DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS UNIVERSIDAD DE CHILE

Speech Processing and Transmission Laboratory

LNFB in Aurora-4

The highest the mismatch, the higher the improvement resulting from LNFB.

Clean training: a 11% reduction in WER

Multi-noise training: a 9% reduction in WER

Multi-condition training (same noise and microphones): a 5% increase in WER

Future research: feature combination

Highly-Reverberant Real Environment database (HRRE)

- Robustness to reverberation is an important problem in ASR
- Several challenges or databases have been generated to address the problem of reverberated speech in ASR: CHiME-2; CHiME-3; CHiME-4; CHiME-5; REVERB; ASpIRE.
- All the reverberated databases that have been employed so far attempt to use real environments, use simulated impulse responses or, in most cases, also include additive noise.

Highly-Reverberant Real Environment database (HRRE)

- Surprisingly, the response of the ASR technologies to RT and speaker-microphone distance has not been addressed methodologically and independently of the additive noise yet.

- There has not been a suitable database for this purpose. HRRE is a response to this need: controlled reverberant environment with several values for the speaker-microphone distance.

- We are covering a wide range of potential applications that span all over from HRI applications, meeting rooms, smart houses to close-talk microphone scenarios.

Highly-Reverberant Real Environment database (HRRE)

- To generate the data for the test set, we re-recorded the original clean test data from the Aurora-4 database (i.e. 330 utterances recorded with the Sennheiser microphone) in a reverberation chamber considering different speaker-microphone distances and RTs.

- The reverberation chamber has an internal surface area of 100 m2, a volume of 63 m3 and an RTmid equal to three seconds. Four reverberant conditions were generated by adding sound-absorbing materials in the reflecting surfaces of the chamber.

Highly-Reverberant Real Environment database (HRRE)

RT=1.77 sec

RT=1.27 sec

RT=0.84 sec

RT=0.47 sec

DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
UNIVERSIDAD DE CHILE

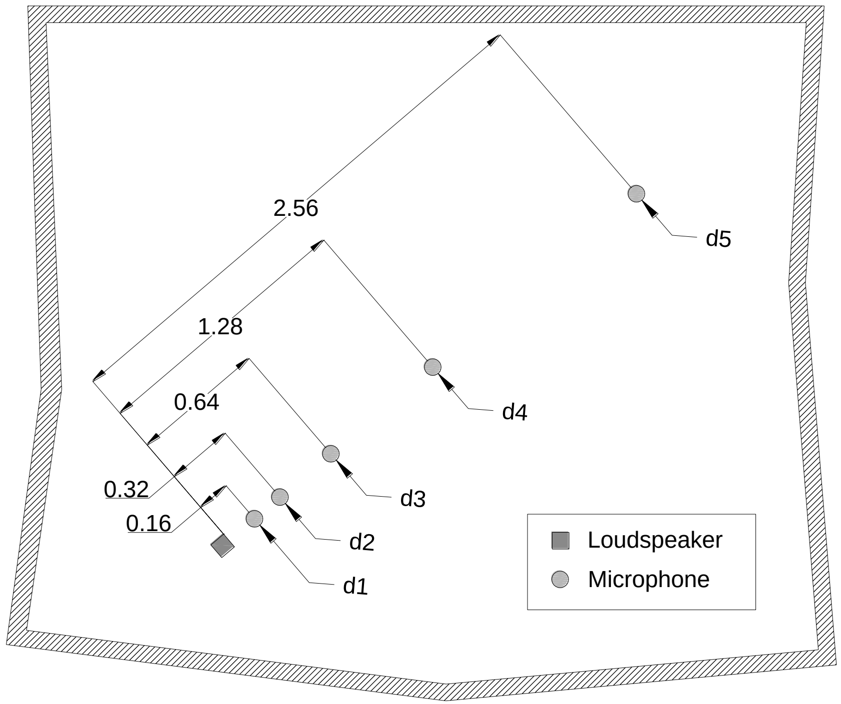
Speech Processing and Transmission Laboratory

Highly-Reverberant Real Environment database (HRRE)

- The loudspeaker-microphone distances were selected as follows. The longest distance was set to 2.56 m. Then, the distance was reduced four times by factors of two ultimately reaching 0.16 m.

- Following this procedure, the selected distances were: 0.16 m, 0.32 m, 0.64 m, 1.28 m and 2.56 m.

Highly-Reverberant Real Environment database (HRRE)

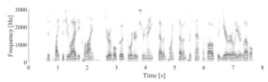


Recording scheme of distances used in the reverberation chamber. The selected loudspeaker-microphone distances were: d_1=0.16 m, d_2=0.32 m, d_3=0.64 m, d_4=1.28 m and d_5=2.56 m.

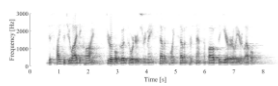
fcfm DEPARTAMENTO DE INGENIERIA ELECTRICA FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS UNIVERSIDAD DE CHILE

Speech Processing and Transmission Laboratory

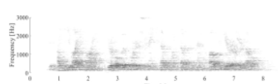
Highly-Reverberant Real Environment database (HRRE)



Clean



RT=1.77 sec
spk-mic=16cm



RT=1.77 sec
spk-mic=2.56m

fcfm DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS UNIVERSIDAD DE CHILE

Speech Processing and Transmission Laboratory

Highly-Reverberant Real Environment database (HRRE)

- Further information and details can be found in the following paper:

J. P. Escudero, V. Poblete, J. Novoa, J. Wuth, J. Fredes, R. Mahu, R. Stern and N. Becerra Yoma, "Highly-Reverberant Real Environment database: HRRE," ArXiv pre-prints 1801.09651, 2018.

Voice-based human-robot interaction

In his book "Usability Engineering", Jakob Nielsen mentioned that "Software developed in recent years has been devoting an average of 48% of the code to the user interface." That was more than 20 years ago.

The more popular the computers became, the more friendly the human-computer interface needed to be.

Voice-based human-robot interaction

The launch of Pepper, the Japanese robot, may indicate that we will be going through the same process in robotics.

In Japan, Pepper is much cheaper than, for instance, PR2, Baxter or NAO.

It was designed to recognize human emotions, and communicate with natural language and gestures.

Voice-based human-robot interaction

It is hard to say if current technology will consolidate Pepper in the market, but what we can say for sure is that the more affordable and popular the robots become, the more friendly human-robot interaction (HRI) needs to be.

And one cannot conceive of a friendly HRI interface without fluid and reliable voice interactions.

Voice-based human-robot interaction

Pepper is not the only one:

Kury, from Mayfield Robotics, a $700 home robot

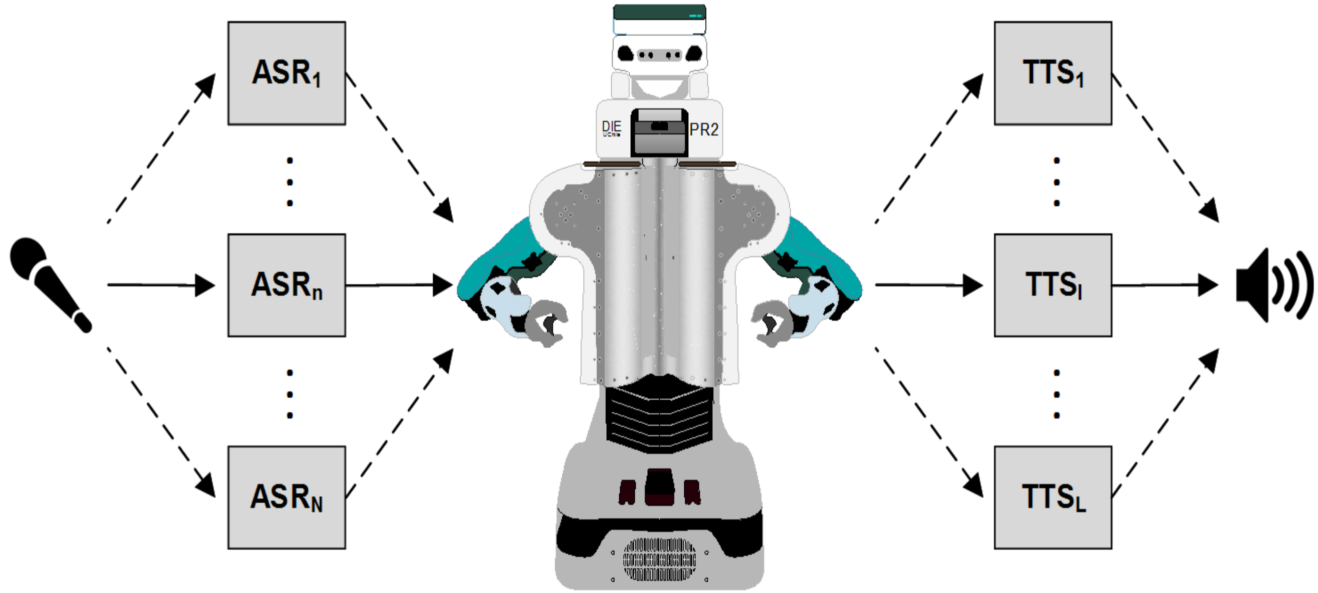
Jibo, another home robot.

DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS UNIVERSIDAD DE CHILE

Speech
Processing and Transmission Laboratory

How do people in robotics see speech?

Speech technology is usually considered as a black box that could be integrated to a robotic system on a surgical basis, rather an integral part of the system itself.



How do people in robotics see speech?

Researcher 1: “Our robot is great. It just needs to talk and hear”

Researcher 2: “Ok! Let´s pick a TTS.   Now let´s pick an ASR”

Researcher 1: “Done! But it does not work”

Researcher 2: “You are right, it does not work”

Researcher 1: “Voila. Now it works. Our baby is ready”

Researcher 2: “Great! Do not move, do not breath. Let´s film it!”

Why does it happen?

- The real potential of speech may be underestimated or overestimated.
- Speech technology is a problem that has already been solved by Google, Apple, Microsoft, etc. Is it really true?
- Some robotic R&D people do not trust speech technology completely, and limited human-robot dialogues are adopted, if any.
- Any thing less than a 100% accuracy solution is not acceptable.

Why does it happen?

- There are many other fundamental problems in robotics.
- There has not been challenges related to voice-based HRI like Robocup.
- It looks like there are not many researchers in robotics that are interested in speech.
- It looks like there are not many researchers in speech that are interested in robotics.

Why does it happen?

As a result, in contrast to computer vision for instance, it looks like speech technology research is underrepresented in robotics.

What about a social robot challenge?

Robotics at LPTV

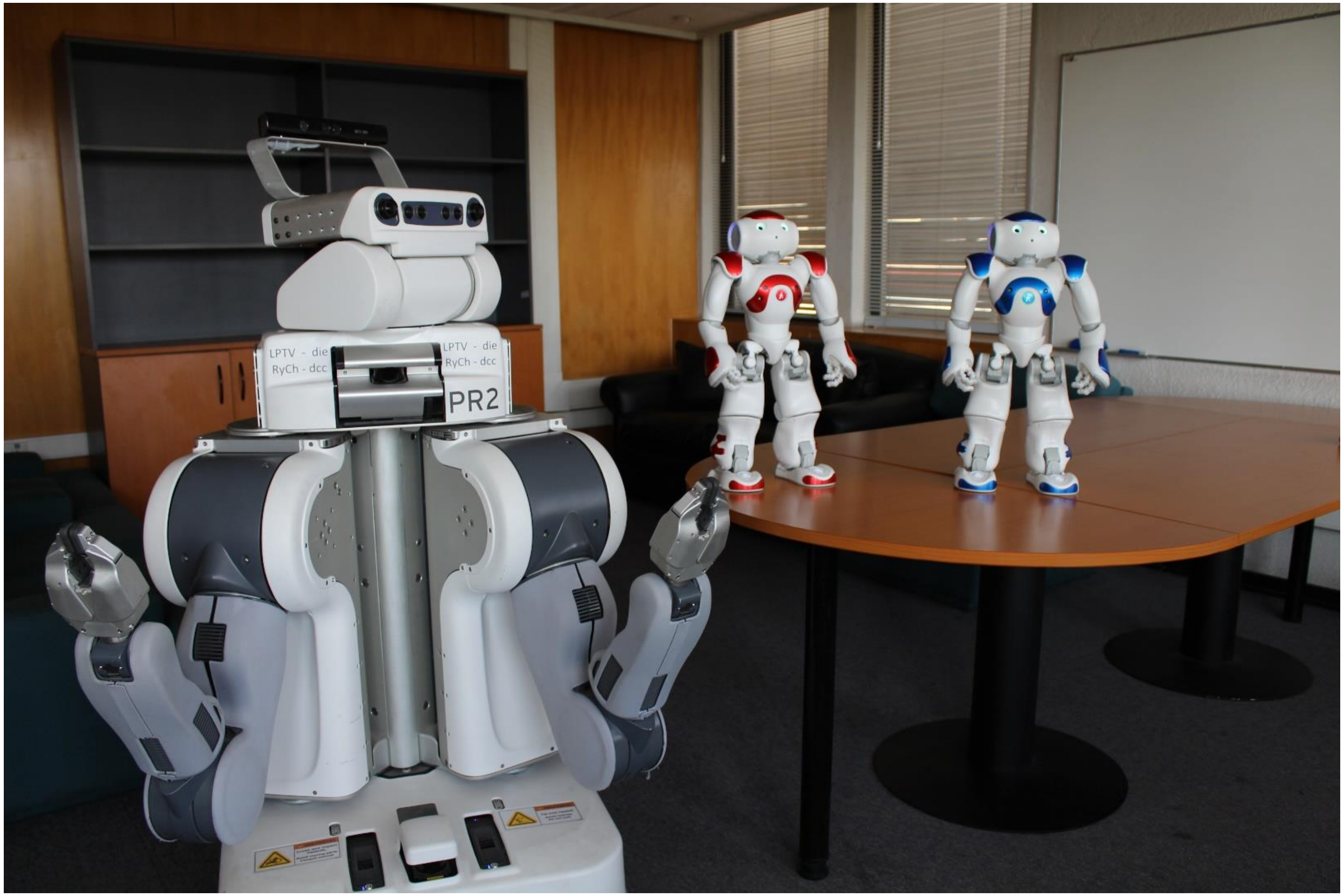
Robotics at LPTV

We are taking care of the last two issues.

We bought a PR2 robot (about US$ 300K) that we call Jarvis.

We also bought two NAO robots, Red and Blue.

Robotics at LPTV

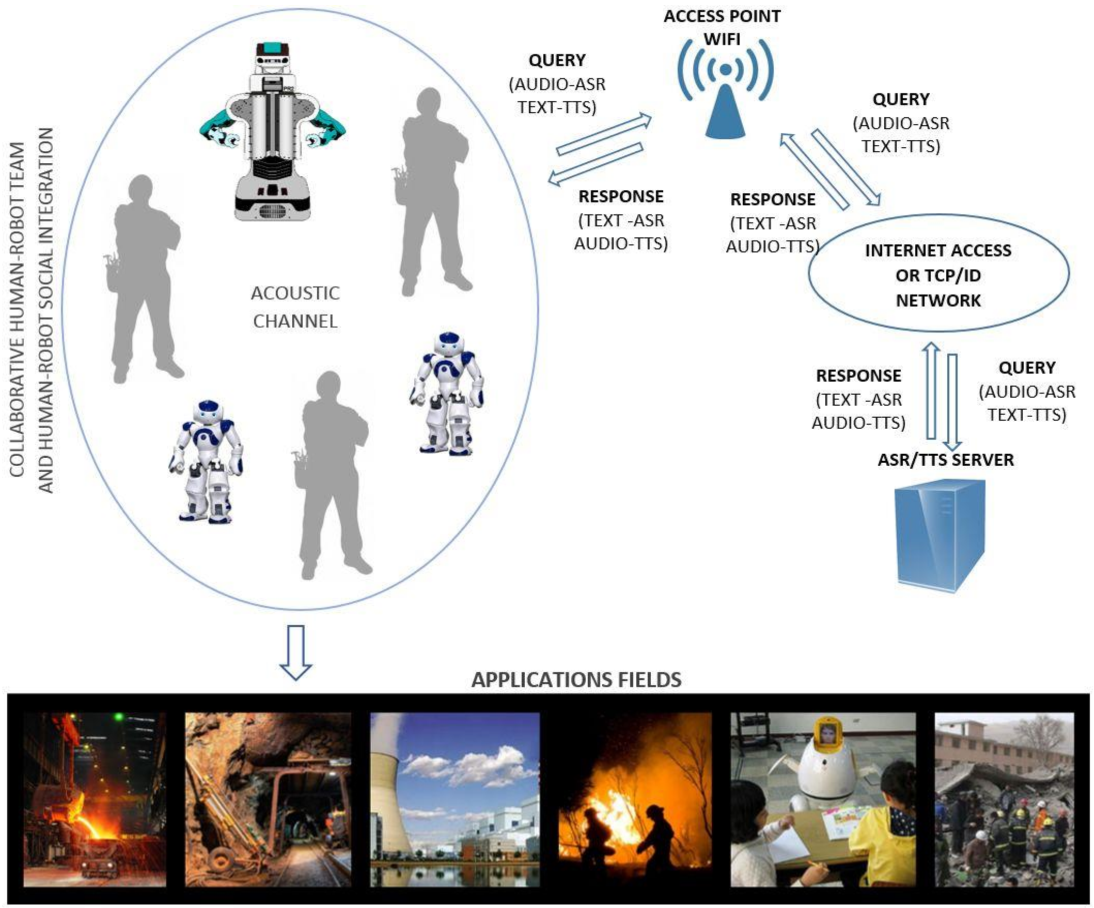


DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
UNIVERSIDAD DE CHILE

Speech Processing and Transmission Laboratory

Robotics at LPTV

Collaborative robotics platform



Some fundamental problems

P1. Minimizing the mismatch between the expectations of human users and the capabilities provided by the available technical solutions.

P2. Defining a general conceptual framework for the interaction of agents.

P3. Modeling and optimizing the spoken dialogue with respect to the robot autonomy, the human operator assistance and the efficiency in task accomplishments.

Some fundamental problems

P4. Improving the robustness of speech recognition to time variant environments.

P5. Improving the robustness of human emotion detection and recognition.

P6. Optimizing the use of the "uncanny valley" effect.

P7. Reproducing human-like emotions with TTS technology.

What about getting inspiration from other fields?

"Spoken language could be the most sophisticated behavior of the most complex organism we know"[1].

From the evolutionary perspective, spoken language has not been the only mode of communication and interactivity [1][2].

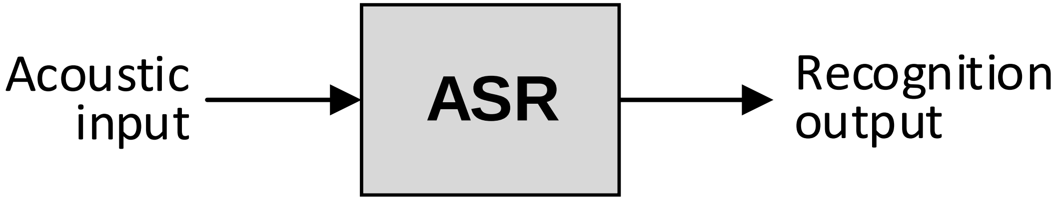
Spoken language is all about context and interactivity rather than "turn-by-turn message passing" [1][2]. See also "enactivism".

[1] Roger K. Moore. "Spoken Language Processing: Time to Look Outside?". Statistical Language and Speech Processing, Springer, pp 21-36, 2014.

[2] Maturana, H.R., Varela, F.J.: The Tree of Knowledge: The Biological Roots of Human Understanding. New Science Library/Shambhala Publications, Boston, (1987).

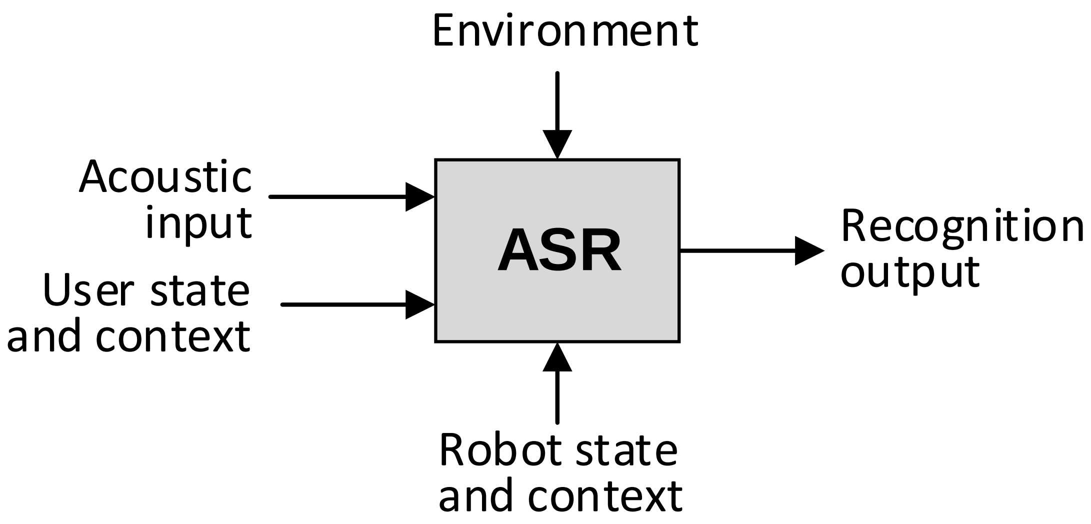
ASR: Are we doing the right thing?

How ASR technology is considered:



ASR: Are we doing the right thing?

How ASR technology should be considered [3]:



[3]Jose Novoa et. al. “DNN-HMM based automatic speech recognition for HRI scenarios.” HRI'18: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction.

DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS UNIVERSIDAD DE CHILE

Speech Processing and Transmission Laboratory

Improving the robustness of speech recognition to time-variant environments

DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
UNIVERSIDAD DE CHILE

Speech
Processing and Transmission Laboratory

Improving the robustness of speech recognition to time-variant environments

This scenario assumes that the robot is sharing the physical space with human beings[4]:

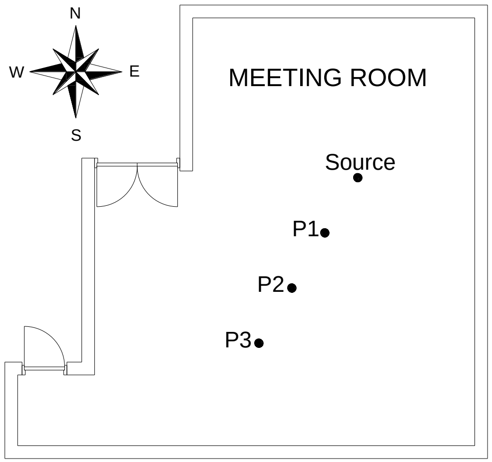
-The displacement speed was chosen so the robot movement would not be too fast for humans.

-The angular speed of the robot head was determined by supposing it is following another person walking at 2km/h or 4 km/h.

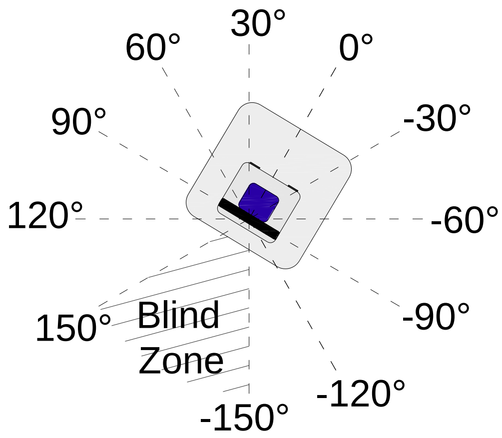
Example: a waiter/waitress in a restaurant.

[4] K. Dautenhahn, M. Walters, S. Woods, K. L. Koay, C. L. Nahaniv, E. A. Sisbot, R. Alami, and T. Siméon, "How may I serve you? A robot companion approaching a seated person in a helping context," in ACM Conference on Human-Robot Interaction (HRI), Salt Lake City, UT, USA: ACM Press, 2006

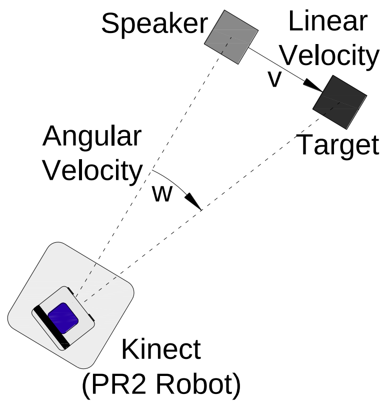
Improving the robustness of speech recognition to time-variant environments



Translational movement of the robot.



Rotational movement of the robot head



Scenario considered

Improving the robustness of speech recognition to time-varying environments

Besides environmental noise, this HRI scenario defines a problem that we called “time-varying channel”.

Ordinary solutions to compensate for channel distortion assume that the channel is linear and time-invariant: an additive constant in the log or cepstral domain. LNFB is an interesting tool in this context.

HRI researchers like to use general purpose API from Google, IBM, Microsoft, etc.

But, we propose to model the acoustic environment as we said.

Improving the robustness of speech recognition to time-varying environments

We went a bit further and recorded an small HRI database for validation purposes with four American English native speakers.

Results were presented at the conference HRI2018 in Chicago (22% acceptation rate):

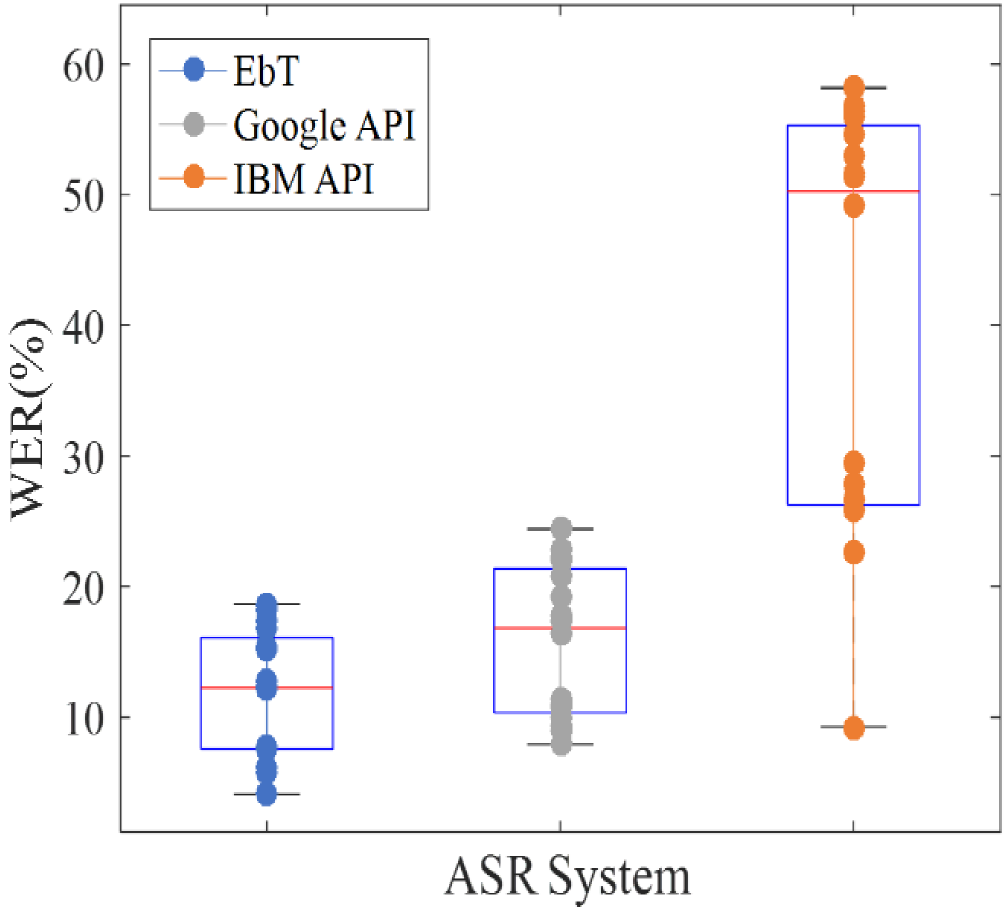
José Novoa, Jorge Wuth, Juan Pablo Escudero, Josué Fredes, Rodrigo Mahu, Néstor Becerra Yoma. HRI '18: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, Chicago, March, 2018.

Improving the robustness of speech recognition to time-varying environments

fcfm DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS UNIVERSIDAD DE CHILE

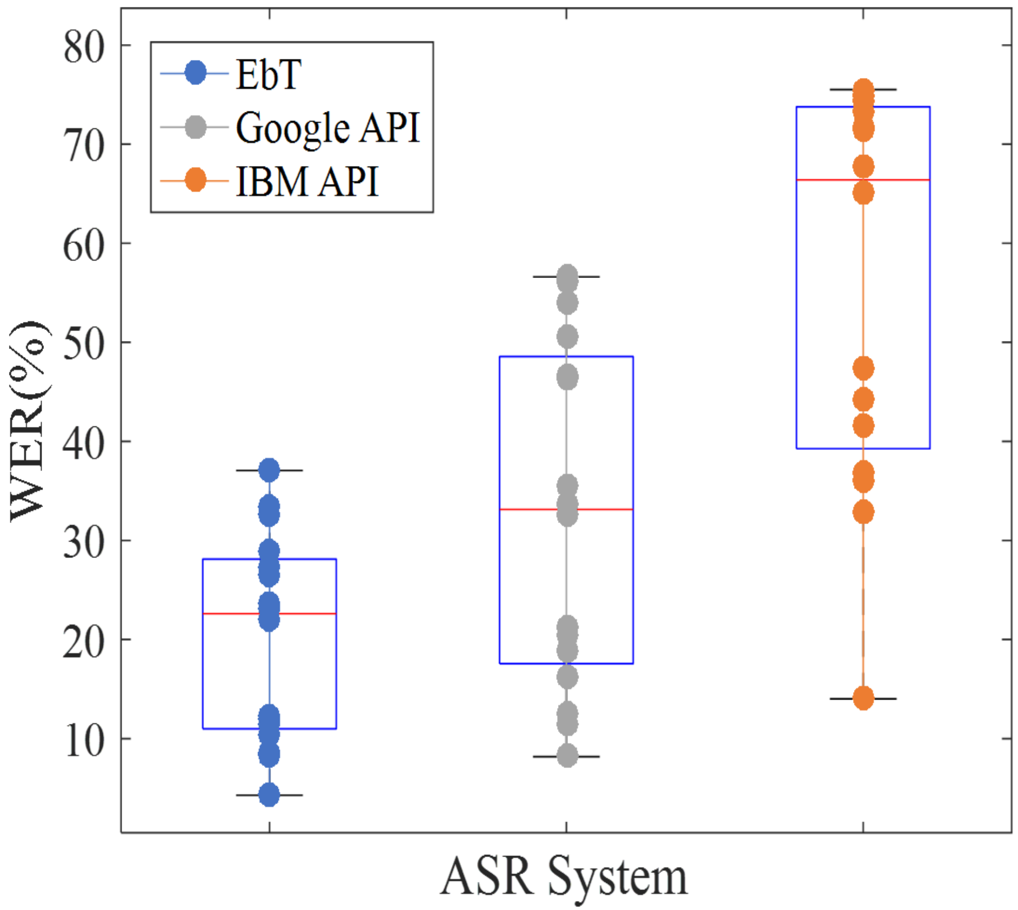
Speech
Processing and Transmission Laboratory

Improving the robustness of speech recognition to time-varying environments



WERs obtained with our EbT (envornment based training) system, Google API and IBM API in all the robot movement conditions, with the playback loudspeaker testing database.

Improving the robustness of speech recognition to time-varying environments



WERs obtained with our EbT (envornment based training) system, Google API and IBM API in all the robot movement conditions, with the American English native speaker testing database.

fcfm DEPARTAMENTO DE INGENIERIA ELECTRICA FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS UNIVERSIDAD DE CHILE

Speech Processing and Transmission Laboratory

TTS: Are we doing the right thing?

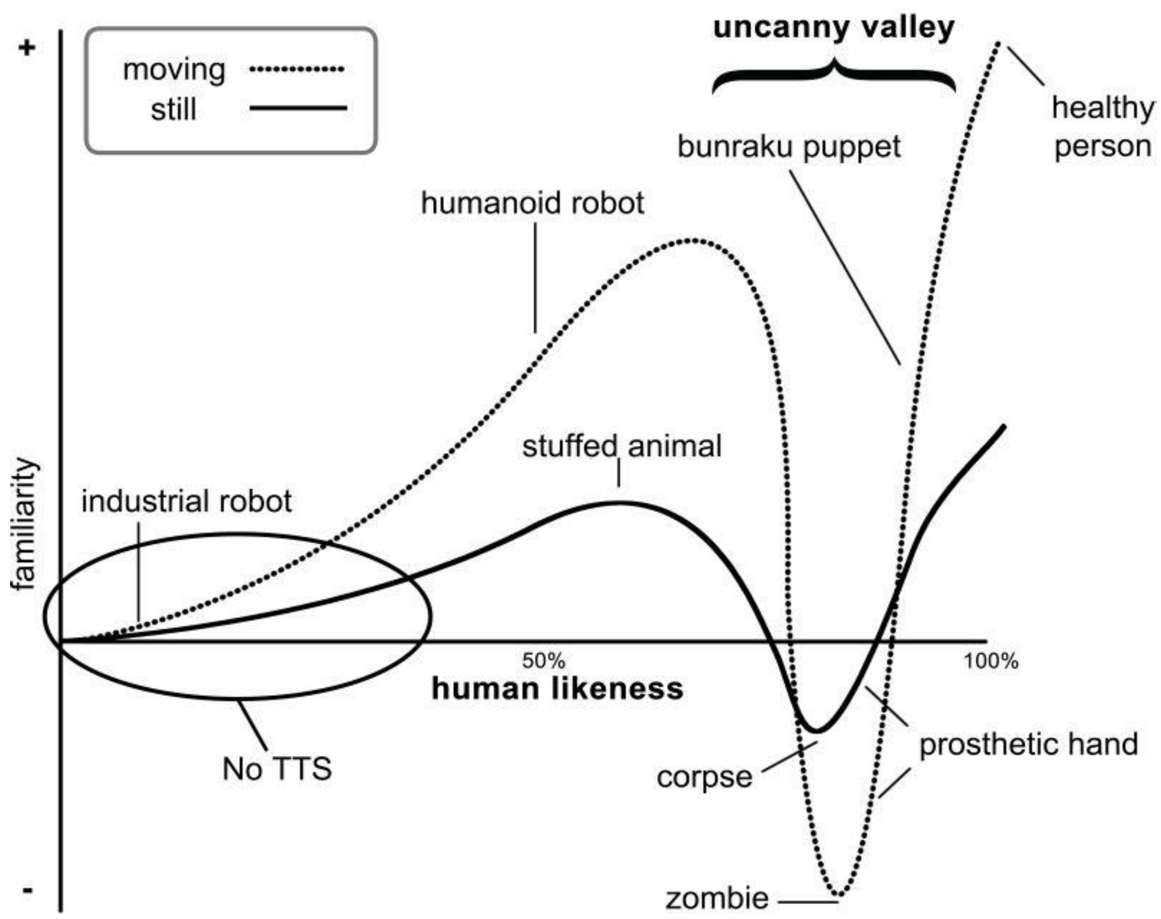
A case to think about:

According to a note in IEEE Spectrum, Kuri home robot does not employs TTS to avoid user´s frustration… Instead, it relies "on a variety of beepy noises and its expressive head and eyes to communicate."

fcfm DEPARTAMENTO DE INGENIERIA ELECTRICA FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS UNIVERSIDAD DE CHILE

Speech Processing and Transmission Laboratory

TTS: Are we doing the right thing?

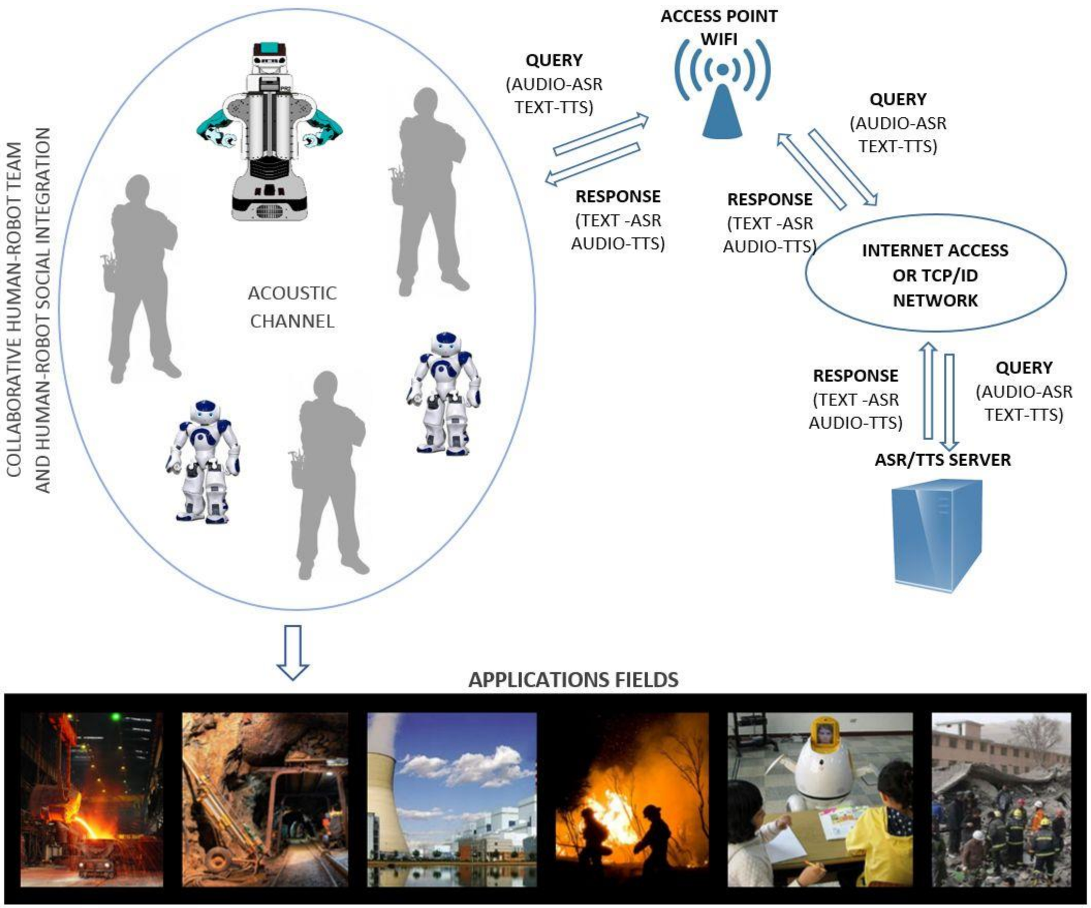


DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
UNIVERSIDAD DE CHILE

Speech
Processing and Transmission Laboratory

TTS and HRI

What role does TTS play in this type of collaborative scenario?



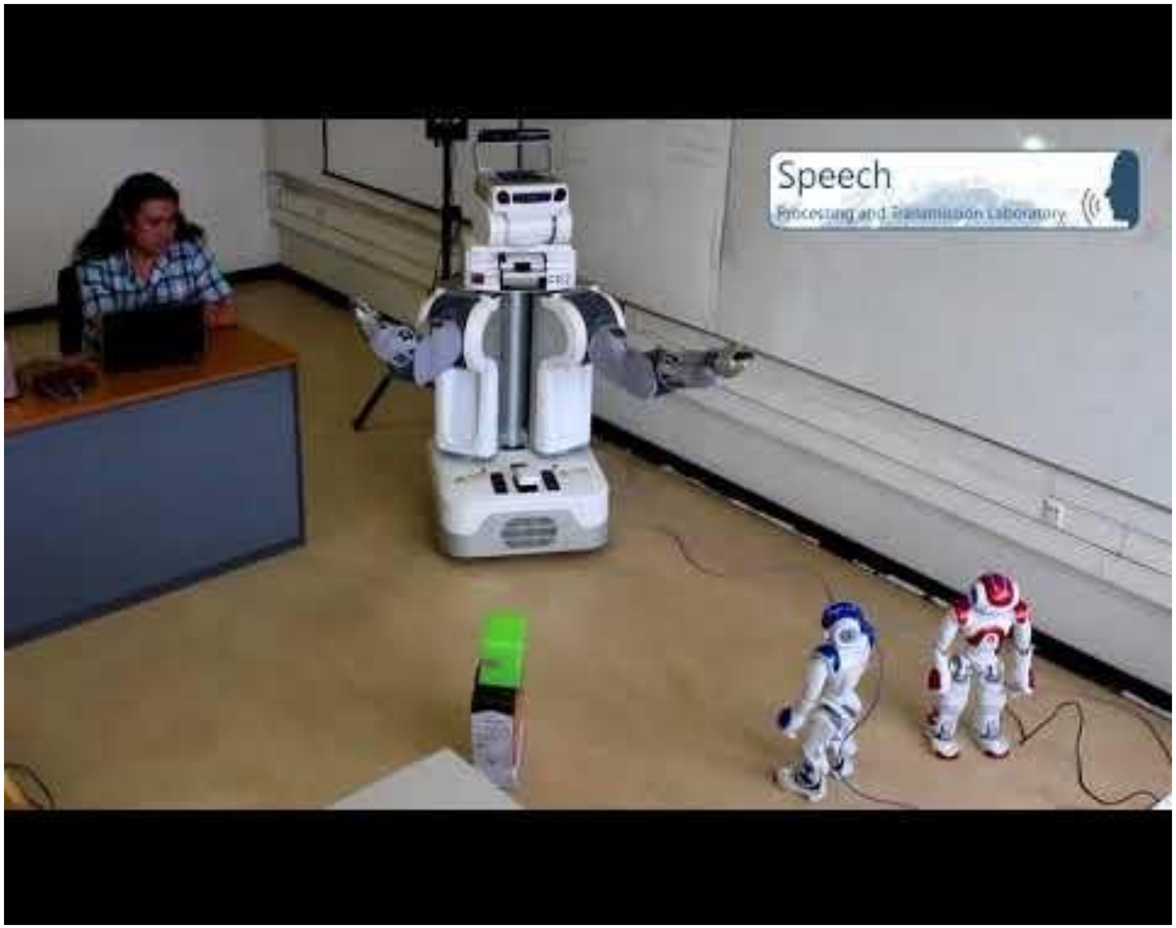
TTS and HRI

Why are you looking at me?

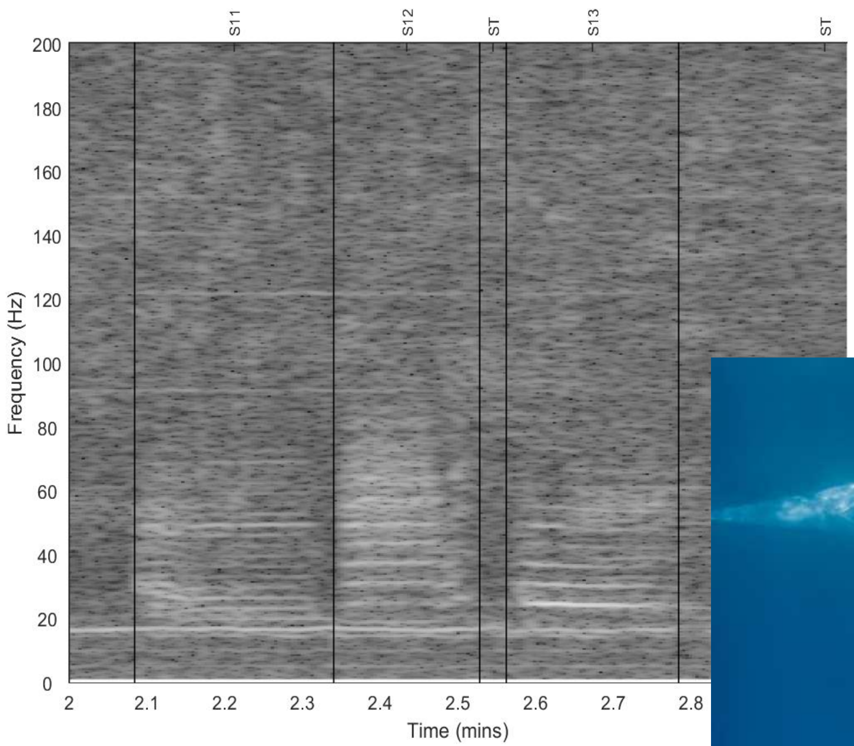
DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
UNIVERSIDAD DE CHILE

Speech Processing and Transmission Laboratory

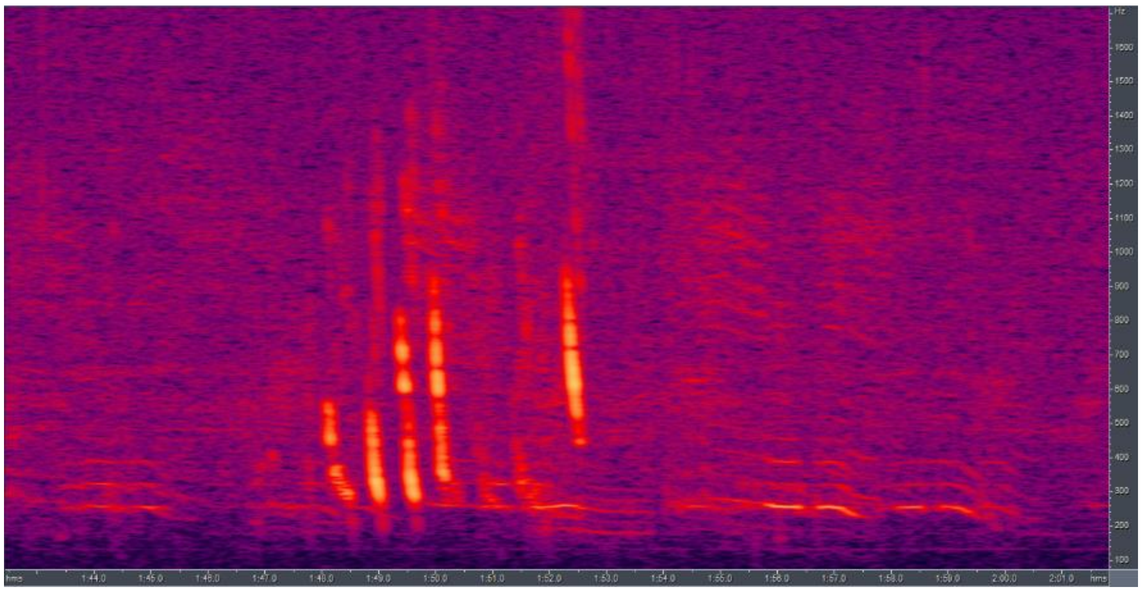
TTS and HRI



Detection and classification of whale vocalizations

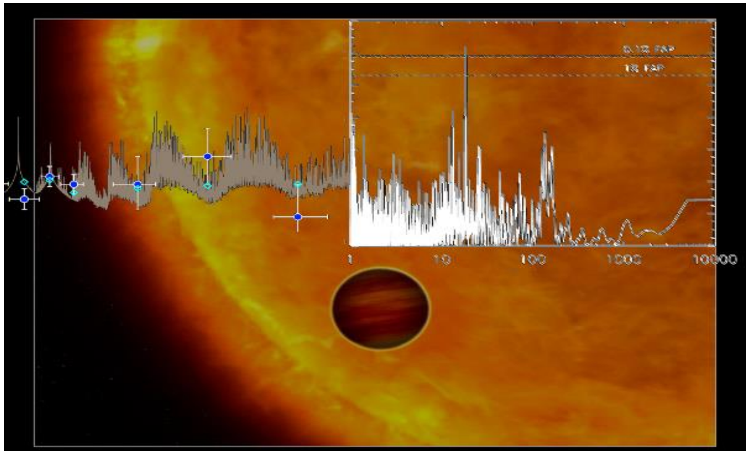


Detection and classification of whale vocalizations

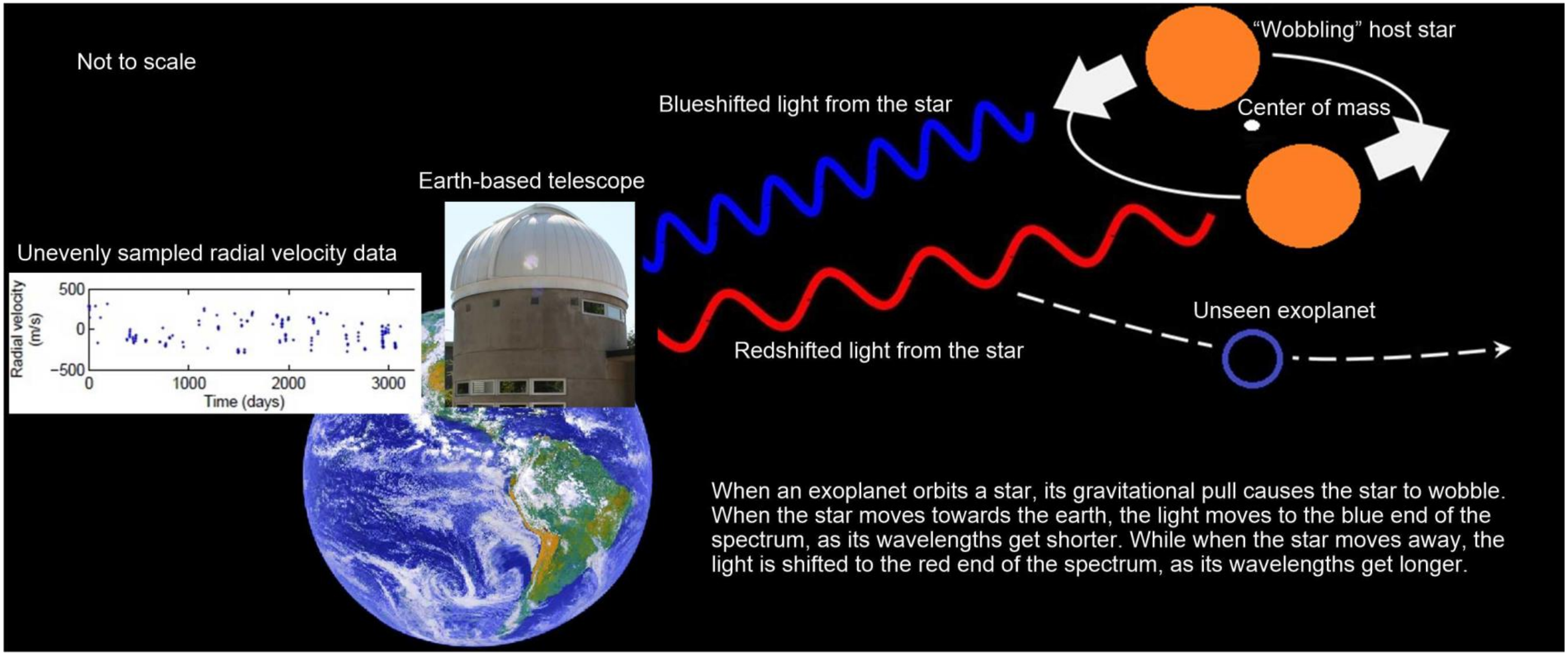


Multidisciplinary research on SP

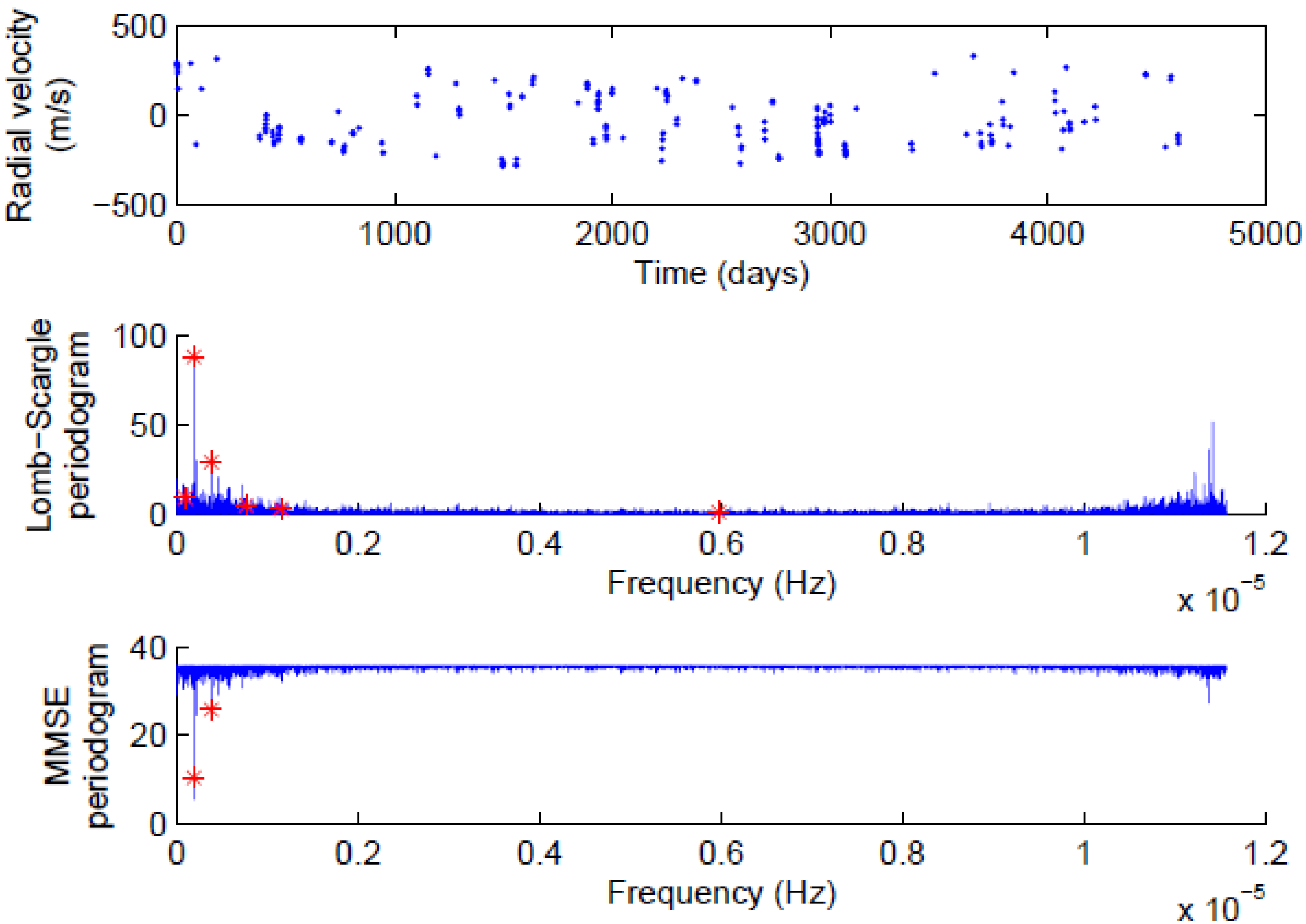
Proyecto Anillo en procesamiento de señales



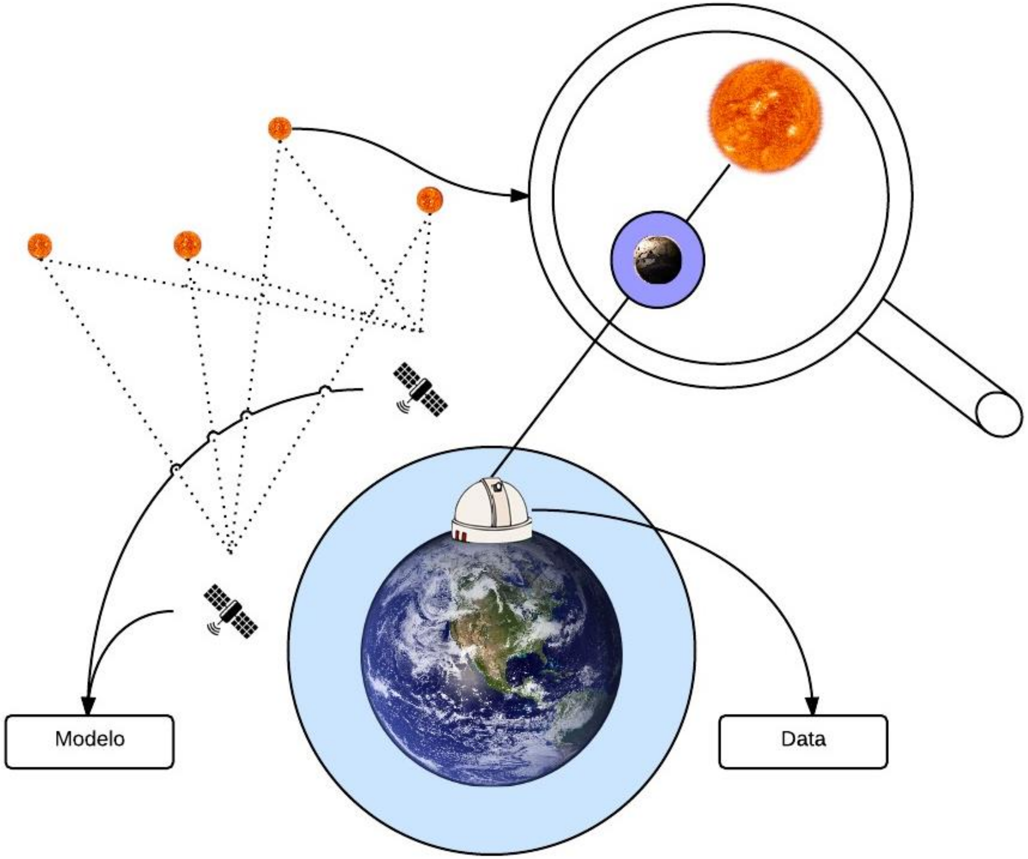
Astronomy: discovering new worlds



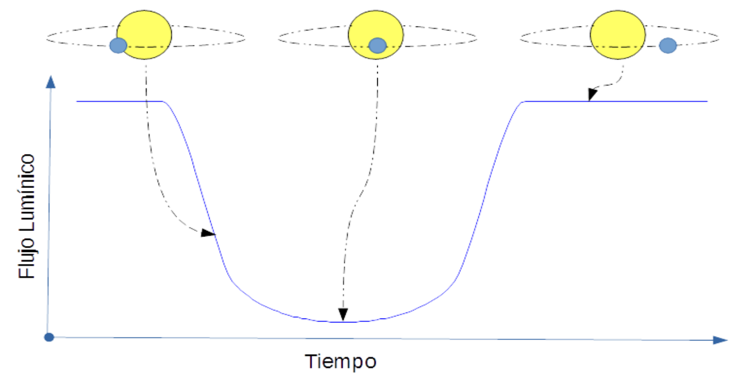
Astronomy: discovering new worlds

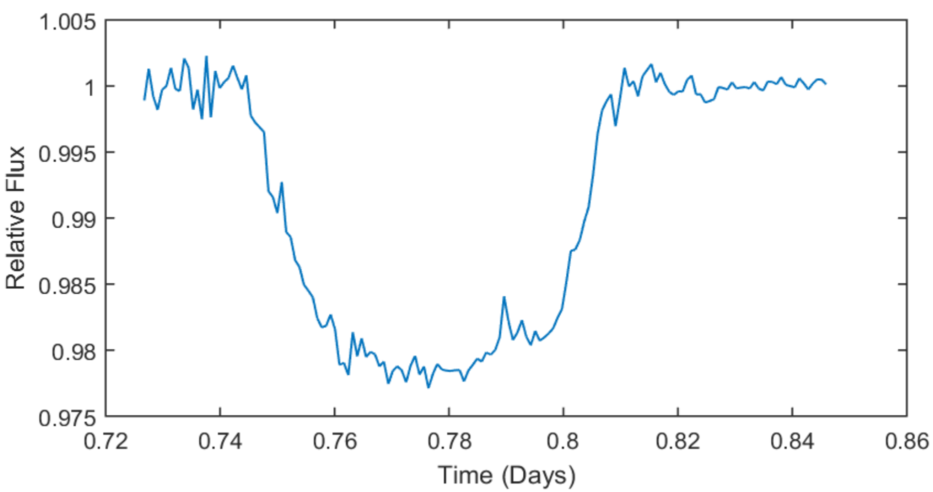


Astronomy: discovering new worlds



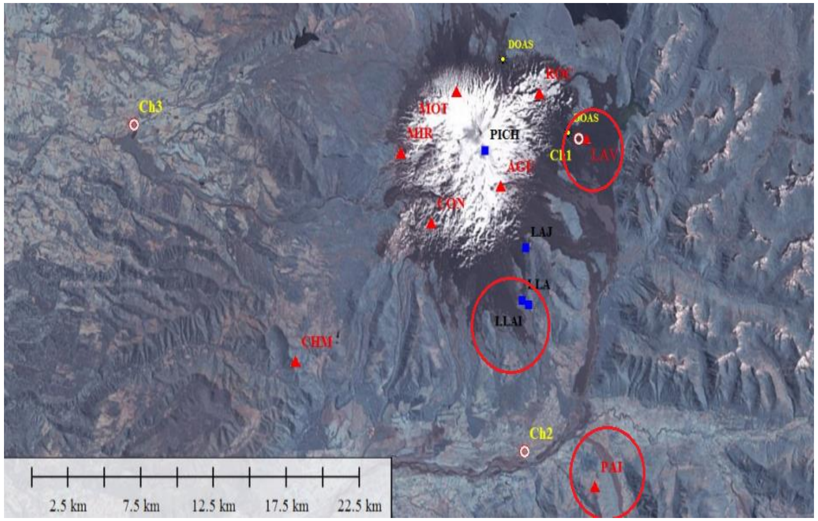
Astronomy: discovering new worlds



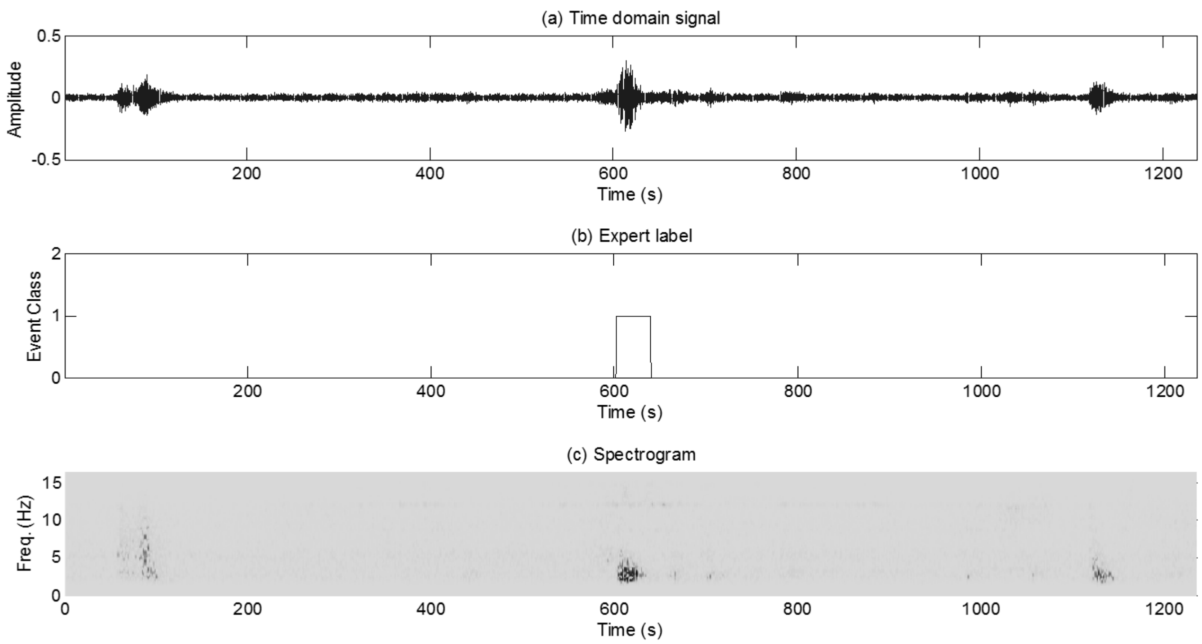


Volcanoes: can we classify their activity?

Llaima Volcano
Chile

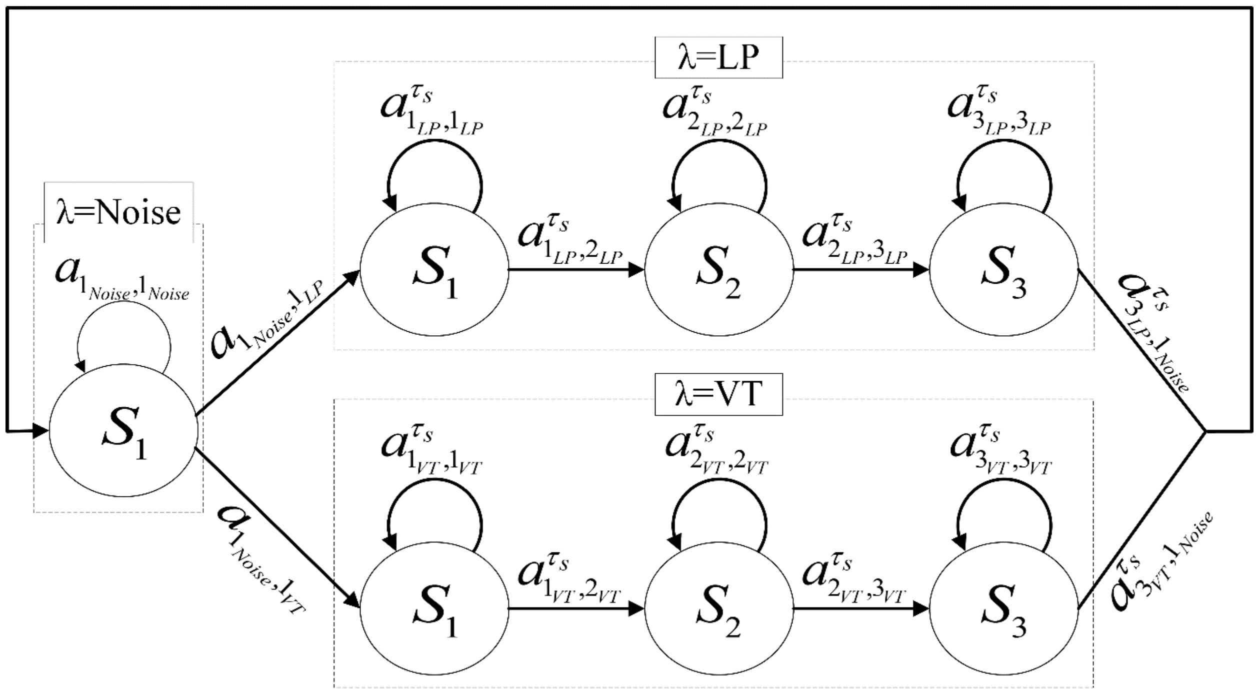


Volcanoes: can we classify their activity?



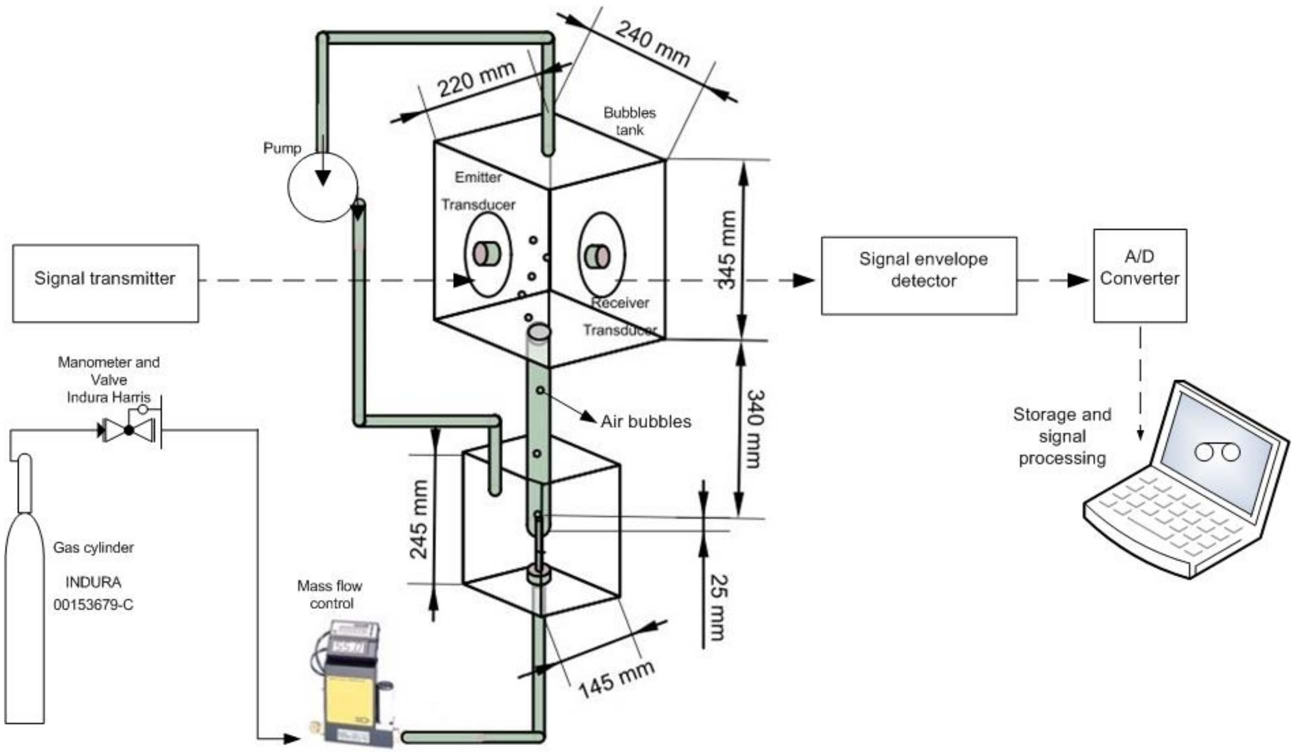
Example of a seismic signal labeled by a volcano expert: (a) time domain signal of a volcano event; (b) labeling according to the volcano expert, where label 0 is noise, label 1 is LP, and label 2 is VT; and, (c) the corresponding spectrogram of the signal.

Volcanoes: can we classify their activity?



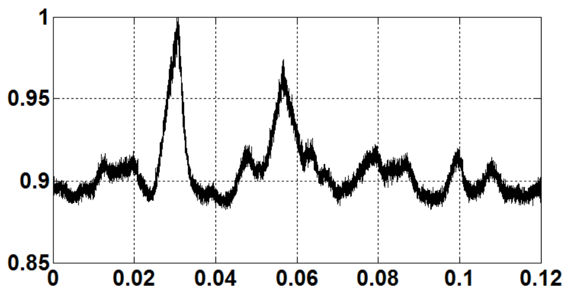
HMM network with state duration modeling for volcano events detection.

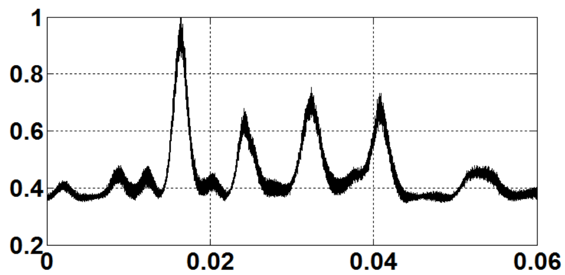
Mining: bubbles are more important than you may think

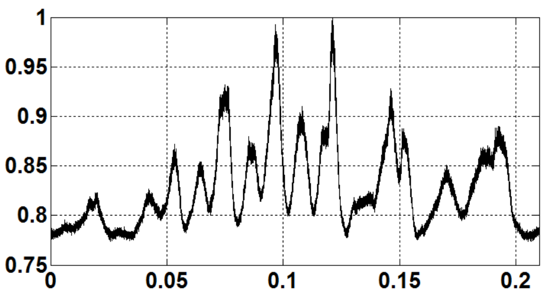


Mining: bubbles are more important than you may think

Signals recorded by the receiver when the bubbles cross the ultrasound beam after envelope detection.







Acknowledgements


Juan Pablo Arias


Alejandro Bassi


Leopoldo Benavides


Carlos Busso


Ignacio Catalán


Felipe Espic

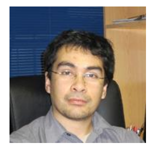

Juan Pablo Escudero


Josué Fredes


Claudio Garretón


Sebastián Guerrero


Alvaro Herrada


Fernando Huenupán


Alvaro Jesam


Salman Khan


Simon King


Rodrigo Mahu


Carlos Molina

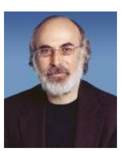

José Novoa


Víctor Poblete


Adolfo Ramírez


Pablo Ravest


Jorge Silva


Richard Stern


Francisca Varela


Hiram Vivanco


Jorge Wuth


Juan Zamorano

CONICYT
COMISIÓN NACIONAL DE INVESTIGACIÓN CIENTÍFICA Y TECNOLÓGICA

Office of Naval Research GLOBAL Science & Technology

fcfm DEPARTAMENTO DE INGENIERIA ELECTRICA
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS UNIVERSIDAD DE CHILE

Speech
Processing and Transmission Laboratory

Thank you

Gracias

Obrigado