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
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.
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
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.


Uncertainty in noise cancelling

Model

\[ x(t) = s(t) + n(t) \]

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

\( S(\omega) \)
\( \hat{S}(\omega) \)
\( N(\omega) \)
\( X(\omega) \)

Noise Removal
Uncertainty in noise cancelling

Model:

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

\[ E[\log(s_m^2(\phi))] = \int_{-\pi}^{\pi} \log(s_m^2(\phi)) \cdot f_\phi(\phi) \cdot d(\phi) \approx \log(E[B_m]) \]

Where

\[ B_m = x_m^2 - n_m^2 \]

\[ E[B_m] = x_m^2 - E[n_m^2] \]
Uncertainty in noise cancelling

Consequently, the uncertainty variance can be defined as:

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

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

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.

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:

\[
b_s(O_t) = \sum_{g=1}^{G} p_g \prod_{n=1}^{N} (2\cdot\pi)^{-0.5} \cdot (Var_{s,g,n})^{-0.5} \cdot e^{-\frac{1}{2} \frac{(O_{t,n}-E_{s,g,n})^2}{Var_{s,g,n}}}.
\]
Then, the expected value of the HMM observation probability is given by:

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

where

\[
V_{tot,s,g,n,t} = \text{Var}_{s,g,n} + \text{Var}(O_{t,n})
\]

\text{Var}(O_{t,n}) \quad \text{is the uncertainty variance.}

\text{Var}_{s,g,n} \quad \text{is the HMM variance}
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:

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

$X$ 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 $\lambda$ is the scaling factor.

Weighted Viterbi algorithm with DNN

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


10 or even 15 years ago, VoIP was a new paradigm. We developed a method to assess 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.

VoIP

Procedure to generate Sequence of lost packets

Evaluation of the packet-loss model
VoIP

Procedure to generate Sequence of lost packets

![Graph showing PESQ score vs. PL (%) for different codecs and bitrates.]

- G.729 [6.4 kbps]
- G.729 [8 kbps]
- G.729 [11.8 kbps]
- G.726 [16 kbps]
- G.726 [24 kbps]
- G.726 [32 kbps]
- G.726 [40 kbps]
- G.711 [64 kbps]
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.

CAPT in the cloud … but in 2007!

Distance between two words with non-linear alignment of models

\[
D(w_x, w_y) = \frac{1}{K} \sum_{k=1}^{K} d(\lambda^m_x(k), \lambda^m_y(k))
\]

<table>
<thead>
<tr>
<th>Spanish</th>
<th>English</th>
<th>Spanish</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>a_s</td>
<td>Ah</td>
<td>m_s</td>
<td>M</td>
</tr>
<tr>
<td>b_s</td>
<td>B</td>
<td>n_s</td>
<td>N</td>
</tr>
<tr>
<td>ch_s</td>
<td>Ch</td>
<td>o_s</td>
<td>Aa</td>
</tr>
<tr>
<td>d_s</td>
<td>D, Dh</td>
<td>p_s</td>
<td>P</td>
</tr>
<tr>
<td>e_s</td>
<td>Eh</td>
<td>r_s</td>
<td>R</td>
</tr>
<tr>
<td>f_s</td>
<td>F</td>
<td>s_s</td>
<td>S</td>
</tr>
<tr>
<td>g_s</td>
<td>G</td>
<td>t_s</td>
<td>T</td>
</tr>
<tr>
<td>i_s</td>
<td>Ih</td>
<td>u_s</td>
<td>Uh</td>
</tr>
<tr>
<td>j_s</td>
<td>Hh</td>
<td>w_s</td>
<td>W</td>
</tr>
<tr>
<td>k_s</td>
<td>K</td>
<td>x_s</td>
<td>KS</td>
</tr>
<tr>
<td>l_s</td>
<td>L</td>
<td>y_s</td>
<td>Y</td>
</tr>
</tbody>
</table>
CAPT in the cloud … but in 2007!

<table>
<thead>
<tr>
<th>Word</th>
<th>Phoneme decomposition</th>
<th>Phonemes</th>
<th>Alignment</th>
<th>Distance between models</th>
<th>(D(w_1, w_2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(w_1)</td>
<td>s ay ah nt ah st t</td>
<td>s ay ah nt ah st t</td>
<td>s ay ah nt ah st t</td>
<td>s ay ah nt ah st t</td>
<td>s ay ah nt ah st t</td>
</tr>
<tr>
<td>(w_{1-SP})</td>
<td>s s i s e s n s t s i s s s t s</td>
<td>s s i s e s n s t s i s s s t s</td>
<td>s s i s e s n s t s i s s s t s</td>
<td>s s i s e s n s t s i s s s t s</td>
<td>s s i s e s n s t s i s s s t s</td>
</tr>
<tr>
<td>(w_{1-EP})</td>
<td>s ih eh nt ih st t</td>
<td>s ih eh nt ih st t</td>
<td>s ih eh nt ih st t</td>
<td>s ih eh nt ih st t</td>
<td>s ih eh nt ih st t</td>
</tr>
</tbody>
</table>

**English phoneme decomposition**

**Spanish phoneme decomposition using spanish units**

**Phoneme replacement**

**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

\[ WDCM_i = \frac{\sum_{r \in E(w_i, H)} Q(h_r)}{\sum_{i=1}^{N} Q(h_i)} \]

Maximum hypothesis log-likelihood

\[ ML_i = \log \left[ \max_r [Q(h_r) \mid r \in E(w_i, H)] \right] \]

Recognition flag

\[ REC_i = \begin{cases} 1 & \text{if } w_i \subseteq h_i \\ 0 & \text{if } w_i \not\subseteq h_i \end{cases} \]

Difference between maximum and minimum state duration

\[ MmSD_i = \text{MaxSD}(t) - \text{MinSD}(t) \]

Position in the N-best

\[ POS_i = \arg \max_r \{[Q(h_r) \mid r \in E(w_i, H)]\} \]
CAPT in the cloud … but in 2007!

The subjective score given a metric can be estimated as:

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

Multi-classifier fusion at feature level:

\[
D(O) = \arg \max_{C_m} P\left(C_m / \overline{WF}(O)\right) = \arg \max_{C_m} \left\{ \frac{P(\overline{WF}(O) / C_m) \cdot P(C_m)}{P(\overline{WF}(O))} \right\}
\]
CAPT in the cloud … but in 2007!

Multi-classifier fusion at abstract level:

- Observation sequence $O$
- ASR output
- $WF_i(O)$
- Classifier 1 $P(C_n/WF_i(O))$
- Decision $d_{WF_1}$
- Classifier 2 $P(C_n/WF_2(O))$
- Decision $d_{WF_2}$
- Classifier J $P(C_n/WF_J(O))$
- Decision $d_{WF_J}$
- FUSION
- $D(O)$
CAPT in the cloud … but in 2007!
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.


Confidence-based classifier fusion

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

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

BBCM: Bayes-based confidence measure.

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

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:

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

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

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

\[ 10 \times \log_{10}(\frac{\|y[n] + y[n - n_i]\|}{\|y[n] - \beta^n y[n - n_i]\|}) - \delta \]

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

$$LNFB_m = \log(LN_m) = \log \left( \frac{LNNum_m}{LNDen_m} \right)$$
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
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.
Highly-Reverberant Real Environment database (HRRE)

Clean

RT=1.77 sec
spk-mic=16cm

RT=1.77 sec
spk-mic=2.56m
Highly-Reverberant Real Environment database (HRRE)

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

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.
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
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”.


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]:

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

Improving the robustness of speech recognition to time-variant environments

Translational movement of the robot.

Rotational movement of the robot head

Scenario considered

Translation, movement, Rotational movement, Scenario considered

MEETING ROOM

Source

P1

P2

P3

N

S

E

W

0°

-30°

30°

60°

90°

120°

150°

-90°

-60°

-120°

-150°

Blind Zone

Kinect (PR2 Robot)

 Linear Velocity

 Angular Velocity

Speaker

Linear Velocity

Target

w

v

[Image: MEETING ROOM diagram with source points P1, P2, P3, and directions N, S, E, W.]

[Image: Rotational movement diagram with angles 0°, -30°, 30°, 60°, 90°, 120°, 150°, -90°, -60°, -120°, -150°.]

[Image: Scenario considered diagram with a speaker and a Kinect (PR2 Robot) labeled with Linear Velocity and Angular Velocity.]

[Image: Graphics from Processing and Transmission Laboratory of Universidad de Chile.]

Speech Processing and Transmission Laboratory
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):

Improving the robustness of speech recognition to time-varying environments
Improving the robustness of speech recognition to time-varying environments

WERs obtained with our EbT (environment 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 (environment based training) system, Google API and IBM API in all the robot movement conditions, with the American English native speaker testing database.
TTS: Are we doing the right thing?

A case to think about:

According to a note in IEEE Spectrum, Kuri home robot does not employ TTS to avoid user’s frustration… Instead, it relies “on a variety of beepy noises and its expressive head and eyes to communicate.”
TTS: Are we doing the right thing?
What role does TTS play in this type of collaborative scenario?
TTS and HRI

Why are you looking at me?
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

When an exoplanet orbits a star, its gravitational pull causes the star to wobble. When the star moves towards the earth, the light moves to the blue end of the spectrum, as its wavelengths get shorter. While when the star moves away, the light is shifted to the red end of the spectrum, as its wavelengths get longer.
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 event 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
Thank you
Gracias
Obrigado