International
Neural
Network
Society

IJCNN 2013
International Joint Conference on Neural Networks
August 4-9, 2013

IEEE
Computational
Intelligence
Society

Audio-visual fuzzy fusion for robust speech recognition

M. Malcangi*, K. Ouazzane, and P. Patel****

****** Università degli Studi di Milano*
Department of Computer Science
DSP&RTS Research Laboratory
*Milan - Italy***

malcangi@di.unimi.it

** London Metropolitan University, Faculty of Computing London, UK

Why

Improvements of robustness of speech recognition is one of the hottest topics in speech signal processing

- ✓ **Combining audio and visual data to implement an audio-visual speech recognition (AVSR) system demonstrated superior performance compared to audio approach alone**
- ✓ **Fuzzy logic-based data fusion method combines optimally the recognition capabilities of audio and visual recognition systems**

What

Speech technology offers a significant advantage over the existing interface to trigger users to change their behavior

- ✓ **On wireless platforms, the technology has not yet reached a level of performance and usefulness to compel users to use the voice interface.**
- ✓ **This research is ultimately aimed to increase ease of use and productivity by producing not a command recognition technology, but a feature rich, continuous, speaker independent bimodal speech recognition.**

How

To implement a robust speech recognition systems it is necessary to integrate both the auditory and visual abilities that a human being uses to recognize a word uttered by another human being

- ✓ **Human beings can surprisingly distinguish between different shapes of the lips, its movement, and associate the lip movement to a group of phonemes (that is, the viseme class)**

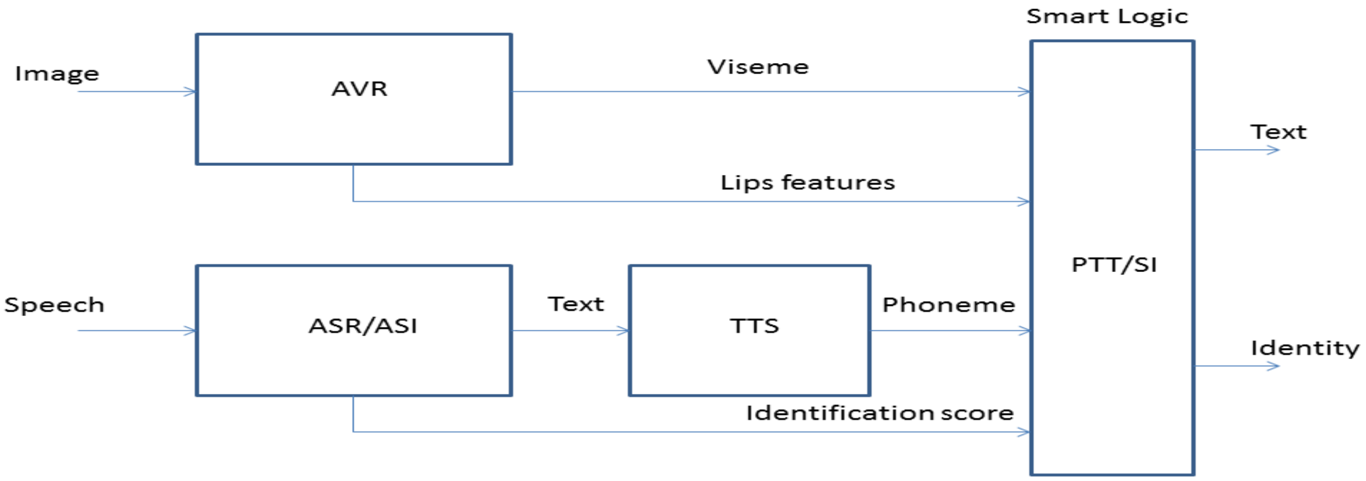
- ✓ **Speech recognition fusion and decision logic needs to be fuzzy so that audio and video information can be optimally combined to infer about the right text corresponding to the uttered word and about the identity of who is uttering**

Key issues

This research aims to establish a framework for the development of bimodal speech recognition and speaker’s identification system based on a fuzzy logic fusion methodology

- ✓ **Speech recognition and speaker identification are executed as standalone processes on isolated uttered words.**
- ✓ **Visual recognition is also executed as a standalone process, focusing solely on lips’ movements during utterance.**
- ✓ **Fuzzy logic executes data fusion recognition at a higher level, keeping the system complexity low while boosting the performance.**

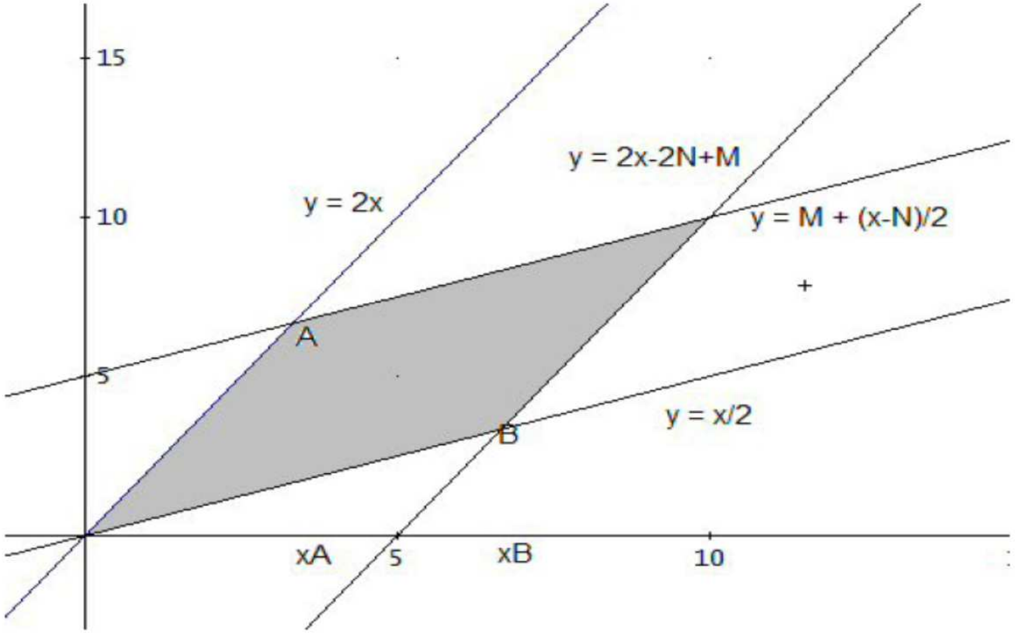
System framework



The system framework consists of four main subsystems:

- ✓ DTW-based automatic speech recognition/speaker identification (ASR/ASI) subsystem
- ✓ Text-to-phoneme transcription (TTP) subsystem
- ✓ Viseme recognition (VR) subsystem
- ✓ Phoneme-to-text transcription/speaker identification (PTT/SI) smart logic-based subsystem

Audio speech recognition (ASR)



The DTW-based ASR subsystem is tuned for speaker-independent isolated word recognition. The uttered word is end-pointed for feature extraction, then passed through a DTW pattern alignment algorithm, and matched on a set of reference words (templates) using Euclidean distance measurement method for similarity evaluation.

Automatic text-to-phoneme transcription

Text-to-phoneme transcription is needed after a text representation of the utterance.

The phoneme sequence needs to be compared to the viseme sequence.

Text-to-phoneme transcription subsystem consists of a rule-based algorithm capable to transcribe a text word into its phonetic sequence.

```
!(P)!=/1p/1i/
(P)!=/1p/4h/4-/
(PA)STE=/1p/1eI/1j/
!(PHOTO)=/1f/1o/1w/1t/2o/2w/
!(PHYS)=/1f/2I/1z/
(PH)=/1f/
(PPH)=/1f/
(PEOP)=/1p/1i/1p/
!(POE)T=/1p/1o/1E/
(POUR)=/1p/1o/13/
(POW)=/1p/1O/1u/
(PP)=/1p/
```

$$C(A)D = B$$

```
!(PRETT)=/1p/1r\/2I/1t/
(PRO)VE=/1p/1r\/1u/
(PROO)F=/1p/1r\/1u/
(PRO)=/1p/1r\/1o/
(PSEUDO)=/1s/2u:/2d/3o/3w/
(PSYCH)=/1s/2a/2a/3j/1k/
!(PS)=/1s/
!(PT)=/1t/
CEI(PT)=/1t/
(PUT)!=/1p/1U/1t/4-/
!(P)=/1p/2H_f/
(P)=/1p/
```

Audio speaker identification (ASI)

$$RMS_j = \sqrt{\frac{1}{N}\sum_{m=0}^{N-1} s_j{}^2(m)}$$

$$ZCR_j = \sum_{m=0}^{N-1} 0.5\left|sign(s_j(m)) - sign(s_j(m-1)\right|$$

$$AC_j = \sum_{i=1}^{N}\sum_{j=1}^{N+1-i} s_j(j)s_j(i+j-1)$$

$$CLPC_j = a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k}$$

- ➢ Using these features, the method scores the person’s identity.
- ➢ The distance measurement of the dynamic time warping—*k*-nearest neighbor (DTW-KNN) algorithm is applied.
- ➢ The cost function is computed using Euclidean distance and the KNN algorithm is used for *k* minimal distance matching.

The purpose of the speaker identification is two-fold:

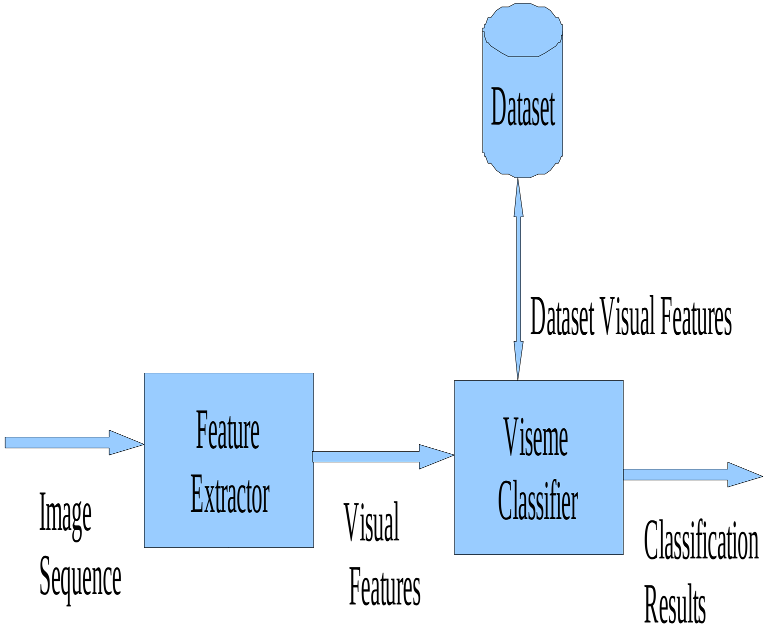
- to improve security level
- to switch automatically to a speaker dependent database that helps to improve speech recognition itself.

Viseme recognition

The viseme recognition subsystem processes the visual representation of the uttered speech and then recognizes the related sequence of visemes.

In visual speech reading, the alphabet detection process involves three main steps:

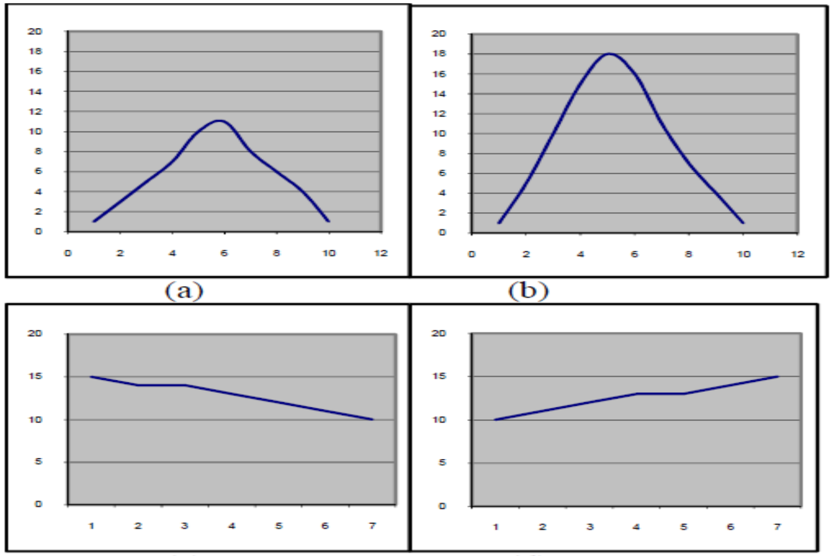
- *Manually extracting visual features for system training and for reference.*
- *Extracting visual features automatically from speaker's facial image sequence.*
- *Comparing these automatically extracted visual features with manually extracted features to find the similarity of spoken alphabet with other alphabets in the database.*



Visual features similarity

The visual features considered are the lip height-width, the time taken for utterance and the facial gestures

- Height similarity score (Hs)
- Width similarity score (Ws)
- Time (no of frames) similarity score (Ts)
- Gesture similarity score (Gs)

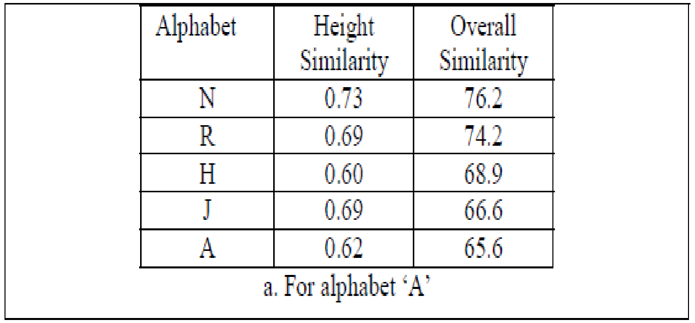


(a) (b)

(c) (d)

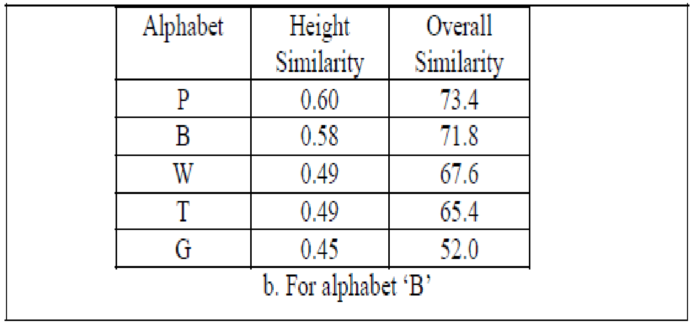
Fig. 4 height variation patterns

HEIGHT SIMILARITY USING UNEQUAL WEIGHTS



Alphabet	Height Similarity	Overall Similarity
N	0.73	76.2
R	0.69	74.2
H	0.60	68.9
J	0.69	66.6
A	0.62	65.6

a. For alphabet ‘A’

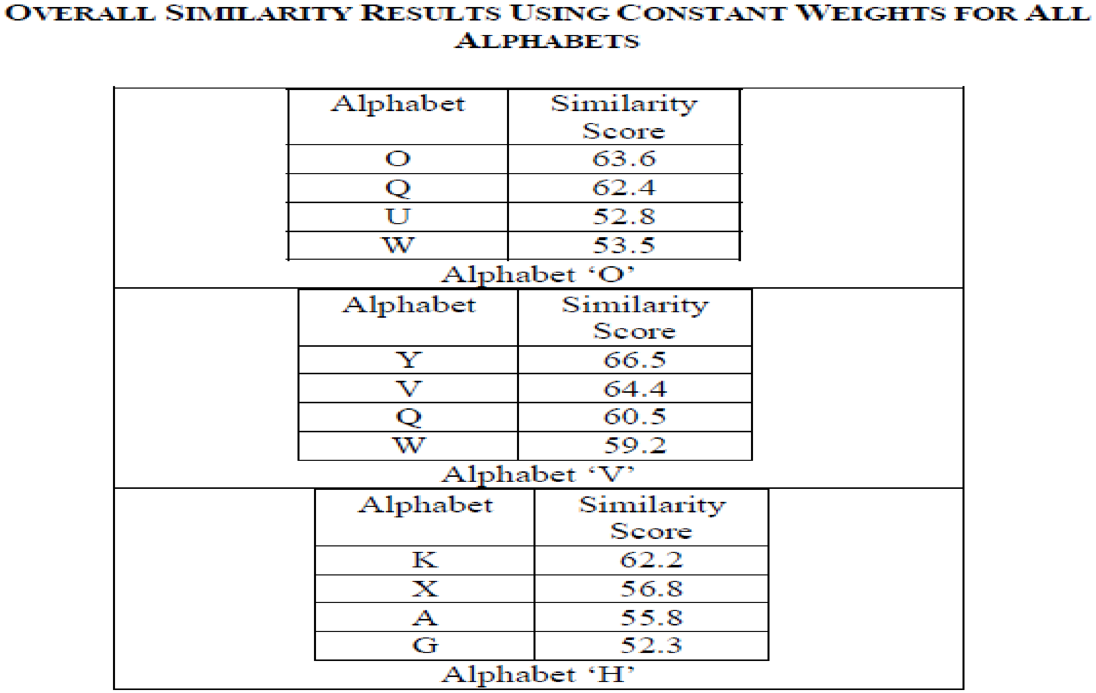


Alphabet	Height Similarity	Overall Similarity
P	0.60	73.4
B	0.58	71.8
W	0.49	67.6
T	0.49	65.4
G	0.45	52.0

b. For alphabet ‘B’

Visual features similarity

An overall similarity is derived by finding the similarity of all four visual features



OVERALL SIMILARITY RESULTS USING CONSTANT WEIGHTS FOR ALL ALPHABETS

Alphabet	Similarity Score
O	63.6
Q	62.4
U	52.8
W	53.5

Alphabet ‘O’

Alphabet	Similarity Score
Y	66.5
V	64.4
Q	60.5
W	59.2

Alphabet ‘V’

Alphabet	Similarity Score
K	62.2
X	56.8
A	55.8
G	52.3

Alphabet ‘H’

As shown in the results, alphabet ‘O’ was correctly recognized as either ‘O’ or ‘Q’

An alphabet ‘V’ was detected as ‘Y’ or ‘V’

Alphabet ‘H’ was not detected correctly

Similarity score was around 60-65%, which is not a very high similarity score

$$Os = 0.35Hs + 0.35Ws + 0.10Ts + 0.20Gs$$

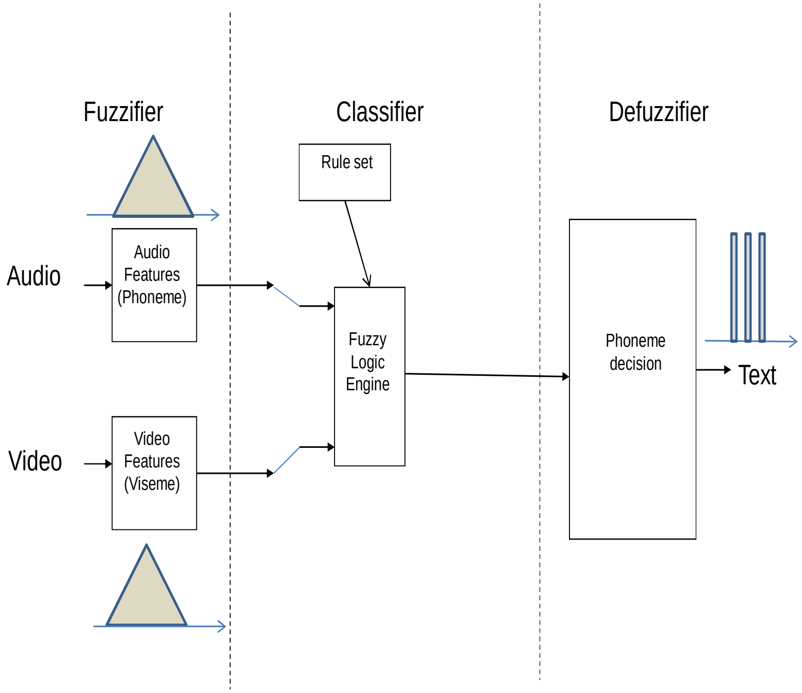
Smart logic for text transcription

A smart logic validates the phoneme sequence using the viseme sequence as reference, corrects phoneme mismatches where they occur and finally generates the text transcription of the uttered word.

The phoneme and the viseme sequences are the two inputs of the smart logic engine tuned to emulate the process that human beings execute when they recognize a word jointly from the sound and from the view of who uttered the word.

If the phoneme and viseme sequences match, then the phoneme sequence is used for phoneme-to-text conversion.

Otherwise the smart logic engine tries to recover phoneme errors where they occur.



Fuzzy modeling

The basic idea is that the phoneme recognition unit and the viseme recognition unit define respectively an audio class A and an image class V so that a set of rules like the following can be applied:

IF Sound IS A_i AND Image IS V_j
THEN Phoneme IS P_k

$P_{i,j,z}$ is the degree matching of A_i and $V_{j.}$ in the phonetic class $P_{z.}$

Fuzzy modeling (cont.)

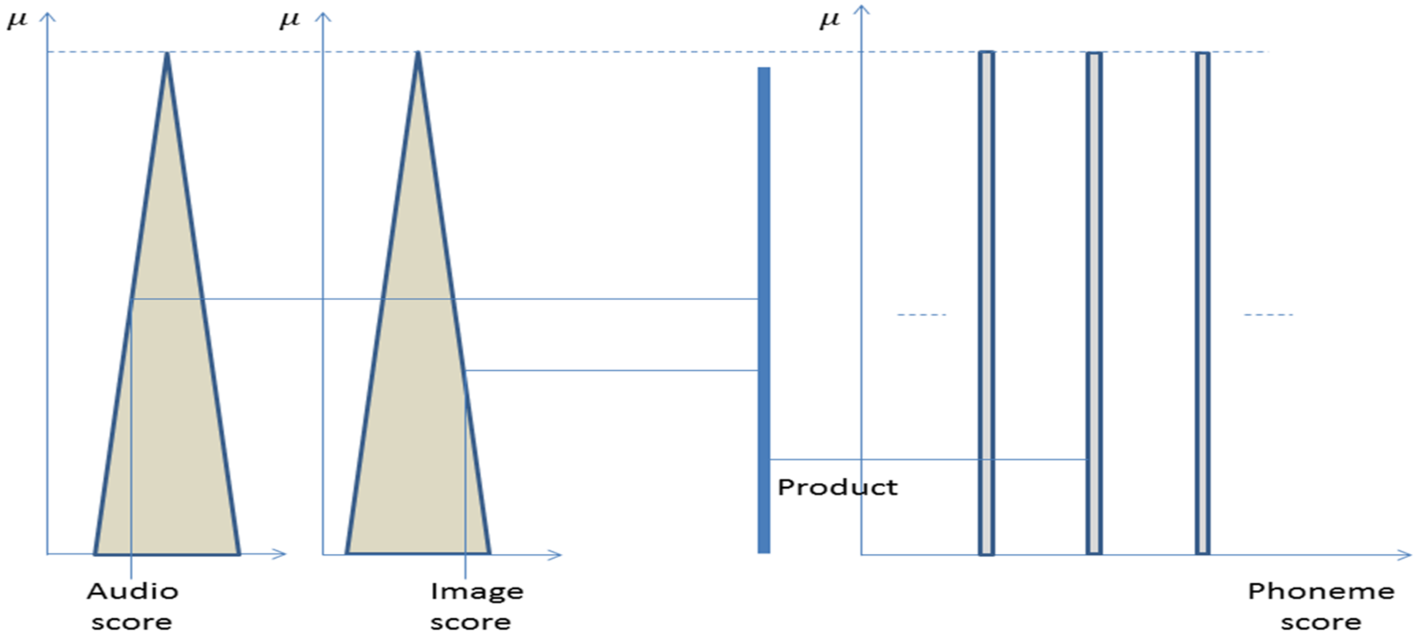
These basic rules can be reinforced with predecessors that consider the previous and the next audio and image too:

```
IF      previous sound    IS  $A_m$  AND
        current sound     IS  $A_i$  AND
        next sound        IS  $A_n$  AND
        previous image    IS  $V_h$  AND
        current image     IS  $V_j$  AND
        next image        IS  $V_k$
THEN    phoneme           IS  $P_z$
```

Fuzzy modeling (cont.)

Each class A and V are fuzzy evaluations executed at audio and image recognition frontend level.

A fuzzy inference engine has been specifically modeled.



Performance

The simulation of the ASR proposed framework shows an average success of 86% in a standalone mode (not fused).

This performance rises to 90% when scored data from visual recognizer are fuzzily fused with the scored data coming from audio recognizer.

The increase in performance for audio recognition can be obtained considering the visual information.

This increase is more evident when audio noise and interferences are heavily masking speech.

In such situations visual data act like a recovery information that keep high the overall recognition rate when the standalone audio recognition module performs bad.

Conclusions

The most important addressed task in this framework concerns data fusion of audio-visual speech information.

Fuzzy logic demonstrates to be an appropriate solution as it enables to emulate the decision logic that human being successfully apply in speech recognition and speaker identification.

Future improvements of this framework should focus on speech analysis and feature extraction at phonetic level.

Conclusions (cont.)

The similarity between audio and visual features at higher level allows for a better integration of the ASR and the AVR.

Segmentation of speech in phonetic units for a better synchronization between audio and visual recognition tasks should also be explored for further improvements of the framework.

Prosodic features such as stress and duration will be extracted from phonetic units (phonemes and allophones) to improve data fusion logic.

Phonetic segmentation of the utterance will be an important improvement to enable unlimited vocabulary speech-to-text capabilities for the system.

More information can be derived from phonetic units that can be correlated with visual information, so that higher speech recognition level will be achieved.

Another long term development is to address the audio-visual natural user interface (AV-NUI).

Thank you for your attention
(any question?)

Mario Malcangi

Università degli Studi di Milano
Department of Computer Science
Via Comelico 39 – 20135 Milano - Italy
DSP&RTS Research Laboratory
(Digital Signal Processing & Real-Time Systems)

Please, address any further question to:

malcangi@di.unimi.it