AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Algorytmy rozpoznawania mowy oparte o kształt i/lub ruch ust - przegląd literatury naukowej z lat 2008-2014

Wojciech Jończyk

Wydział Inżynierii Mechanicznej i Robotyki
Katedra Mechaniki i Wibroakustyki

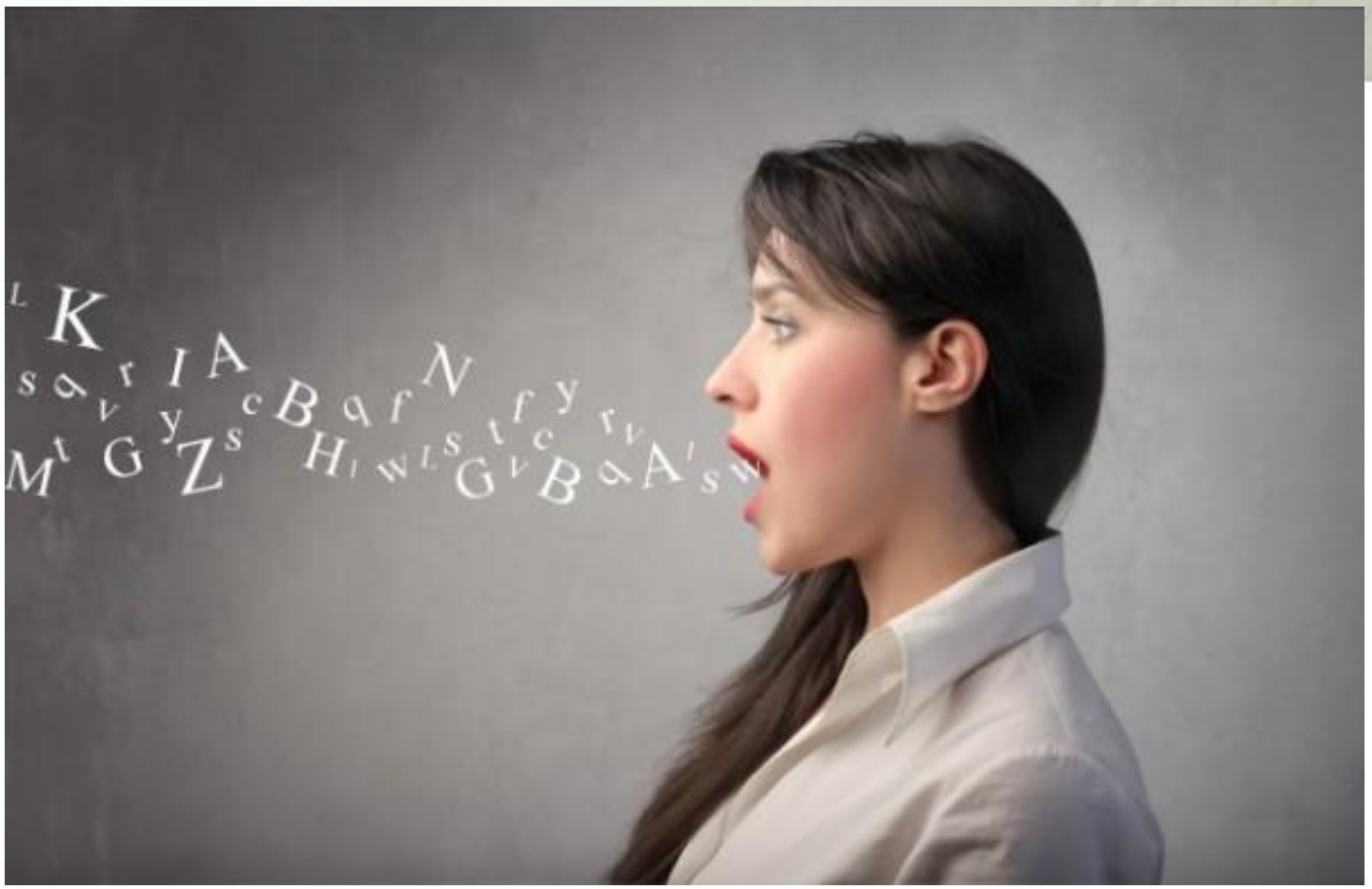
Miejsce i data prezentacji: Kraków, 12.01.2015 r.

Plan prezentacji

- Przybliżenie tematu
- Opis źródeł literatury naukowej
- Przykładowe algorytmy i ich skuteczność
- Podsumowanie

Dlaczego?

- Rozpoznawanie mowy jest jednym z najważniejszych zadań współczesnej informatyki.
- W środowisku o niskim SNR informacje płynące z samego dźwięku nie są wystarczające w identyfikacji i rozróżnianiu mowy i mówców.
- Bimodalizm ludzki percepcji mowy.



Źródło: http://blog.mediaprojectgroup.com/wp-content/uploads/2012/11/18-600x388.jpg

Quentin Summerfield

http://www.york.ac.uk/media/news-and-events/features/summerfield.jpg

Korzyści płynące z analizy

Zapewnia informacje o:

- Źródle dźwięku
- Położeniu artykulatorów
- Segmentach mowy

SPEAKER INDEPENDENT VISUAL-ONLY LANGUAGE IDENTIFICATION

Jacob L Newman and Stephen J Cox

**School of Computing Sciences,
University of East Anglia,
Norwich, UK**

ABSTRACT

'We describe experiments in visual-only language identification

(VLID), in which only lip shape, appearance and motion are used

to determine the language of a spoken utterance.'

VLID

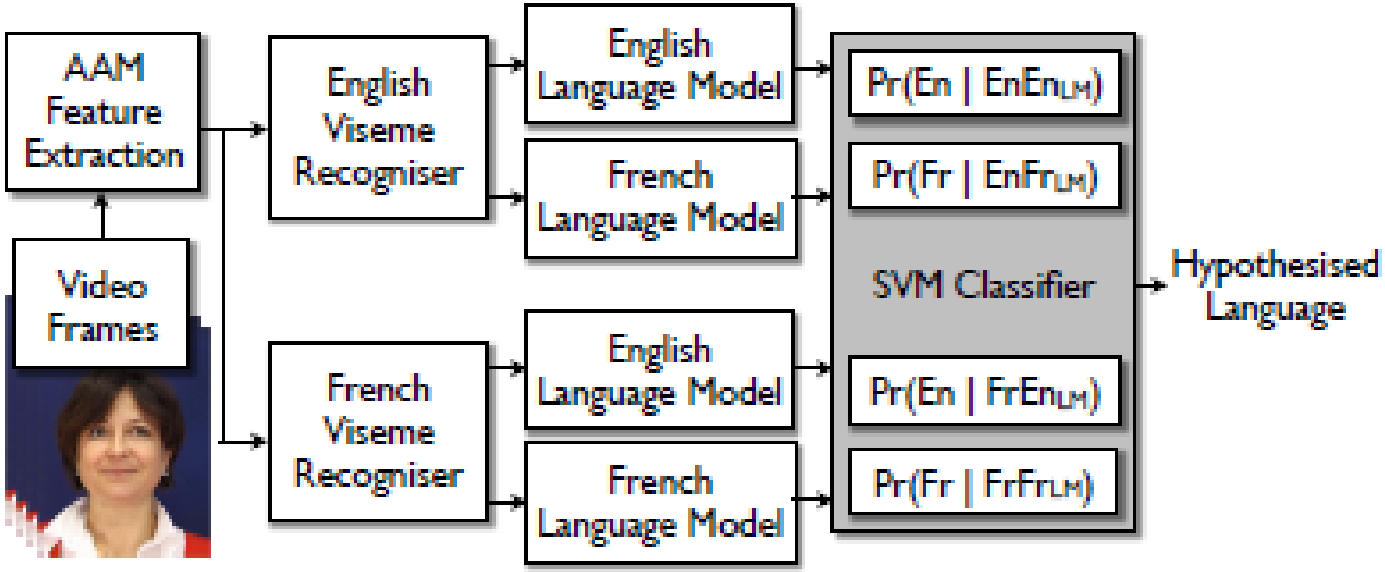


Fig. 1. Visual-only LID System Diagram

Algorytm

- Badano 5 osób płynnie mówiących po angielsku i francusku, 3 z nich uczyły się języków od urodzenia
- Czas trwania testów zróżnicowany, 60, 30, 7, 3 i 1s.
- Rozpoznawano tzw. viseme, czyli charakterystyczne ułożenie ust dla danego fonemu.

Wyniki badań

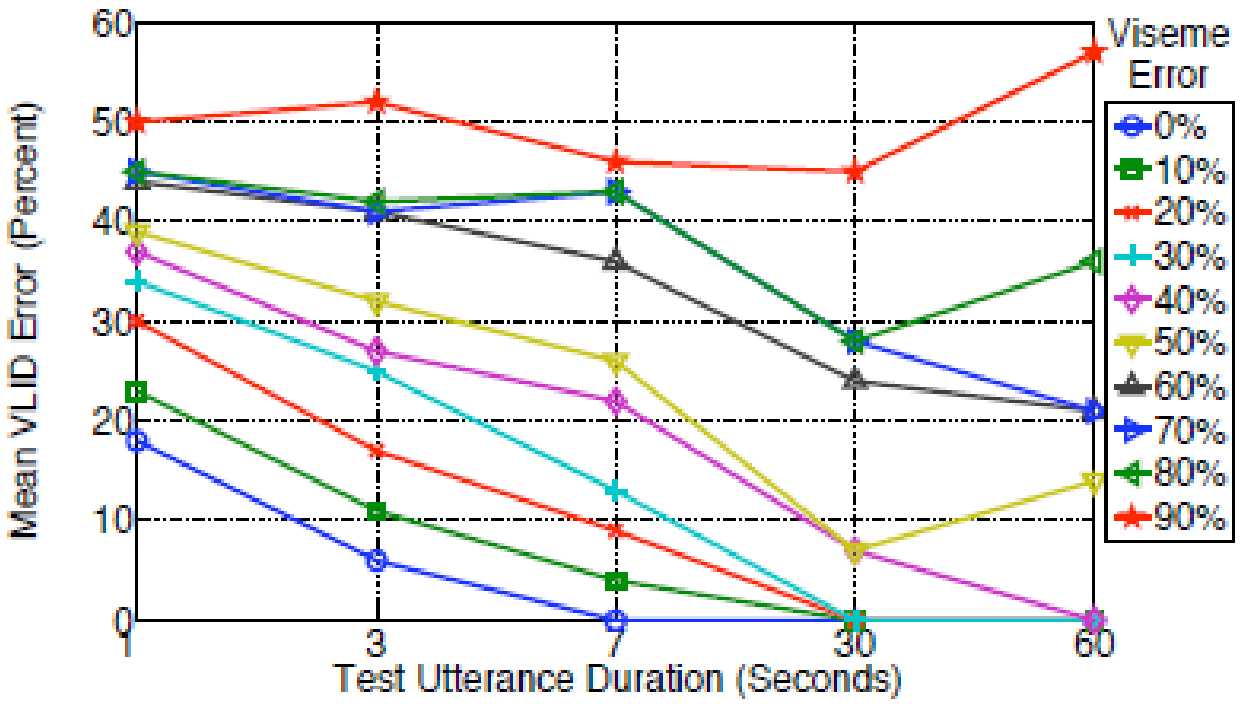


Fig. 3. The effect of viseme accuracy on VLID recognition

Badania wykazały, że nawet na poziomie **40%** złego określenia viseme skuteczność VLAD dla badania o czasie trwania **60s** może wynieść **100%!**

Lip Localization and Viseme Classification for Visual Speech Recognition

Salah Werda, Walid Mahdi and Abdelmajid Ben Hamadou

Abstract

‘The need for an automatic lip-reading system is ever increasing. Infact, today, extraction and reliable analysis of

facial movements make up an important part in many multimedia systems such as videoconference, low

communication systems, lip-reading systems.’

ALiFE (Automatic Lip Feature Extraction)

Metoda ta składa się z 3 kroków:

- Lokalizacja i śledzenie ust

- Wyciąga dokładnie i trafne cechy ludzkiej twarzy

- Wyodrębnione cechy są używane do klasyfikacji i rozpoznania viseme

Schemat

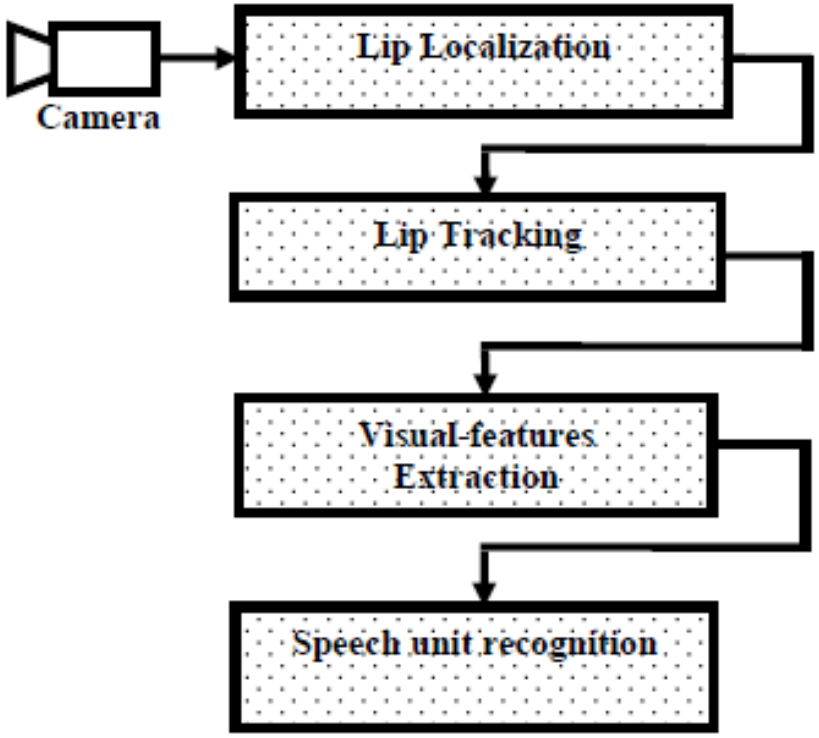


Figure 1. Overview of the complete ALiFE System for speech recognition

1. Lokalizacja ust i wyznaczenie ich POI(Points of Interest)

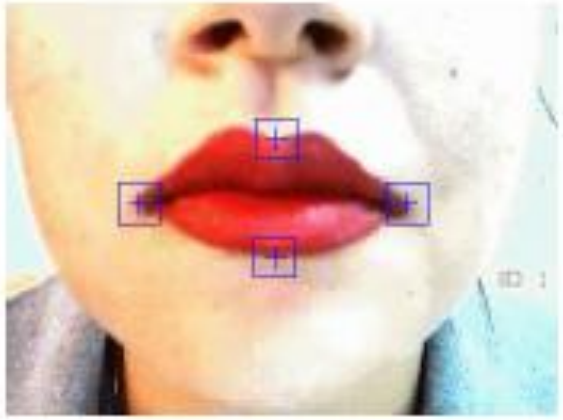


Figure 2. The Different POI for the extraction stapes

Wyznaczano punkty, które głównie odpowiadają
Za możliwą deformację ust

2. Wyznaczanie i obserwacja zmian energii w POI(Points of interest).

W każdym z POI można określić trzy rodzaje energii:

- Energia wewnętrzna

- Energia potencjalna

- Energia ograniczeń nakładana przez daną osobę

Badanie występujących zmian w poszczególnych energiach prowadzi do określenia i wyznaczenia danych ***viseme***, które z kolei odpowiadają odpowiednim fonemom, dzięki czemu możliwa jest analiza i rozpoznanie mowy.

Sposób rejestracji obrazu

Figure 8. Headset of the Labiophone project on a synthesis clone conceived by Ganymédia [9]

Kodowanie Freemana

Metoda ta służy do prostego i szybkiego opisu kierunku ruchu, kontury będą reprezentowane za pomocą sekwencji kroków od pierwszego do ósmego kierunku, które są oznaczane liczbami całkowitymi od 0 do 7.

Kodowanie Freemana

Figure 9. Different directions of Freeman coding

Inny sposób określania cech ust – metoda Dark Area

- Wykorzystuje przestrzeń między górną i dolną wargą do analizy, oceny i przyporządkowania do odpowiedniego fonemu
- Dwa wskaźniki, odchylenie vertical i horizontal

Dark Area

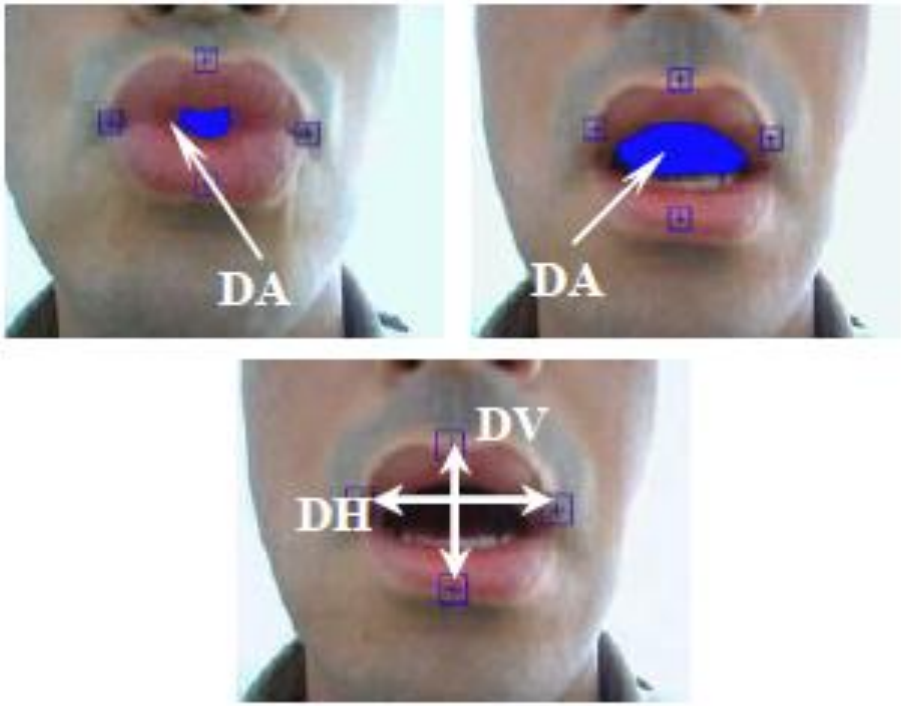


Figure 13. Different descriptors for extraction to the recognition stage

Dark Area

- Analizuje, które piksele reprezentują wnętrze a które zewnętrze ust
- Problemy z określeniem dokładnej granicy miedzy zmianą ośrodka
- Zastosowanie odpowiednich algorytmów minimalizuje błąd złego określenia ciemnej(dark) i nie(non-dark) części

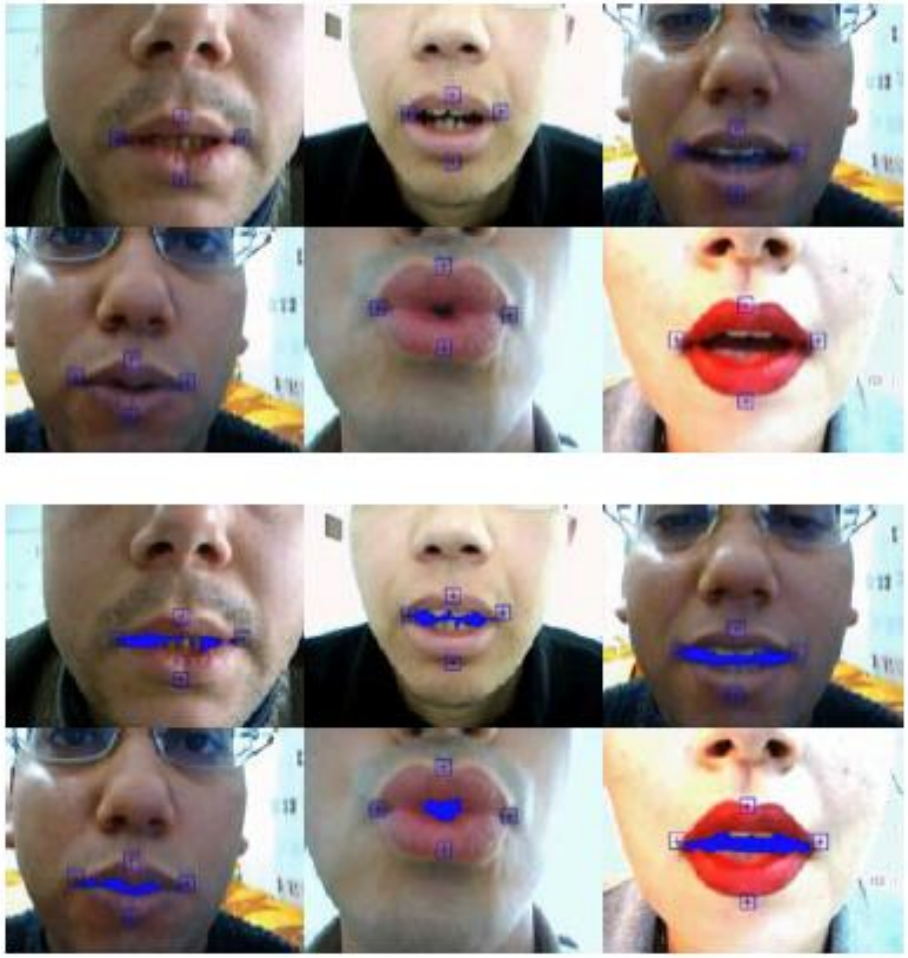


Figure 15. Result of the dark area extrzaction with various speakers under different lighting conditions (From left to right and from top to bottom the images without detection then the same images with the dark area detection)

Viseme corpus

Do stworzenia bazy viseme zaproszono 42 obywateli Francji w różnym wieku i obu płci. Nagrania wykonano z rozdzielczością 0,35 megapiksela i 25 fps.

Następnie zaczęto prowadzić testy i badania.

Wyniki badań

	ba	bi	bou	Recognition Rate
ba	19	7	4	63.33%
bi	4	22	4	73.33%
bou	3	2	25	83.33%

Tab.2: Recognition rate of French Vowel Training Stage

	ba	bi	bou	Recognition Rate
ba	7	4	0	63.64%
bi	3	8	0	72.73%
bou	0	2	9	81.82%

Tab.3: Recognition rate of French Vowel Recognition stage

Automatic Outer Lip Contour Extraction in Facial Images

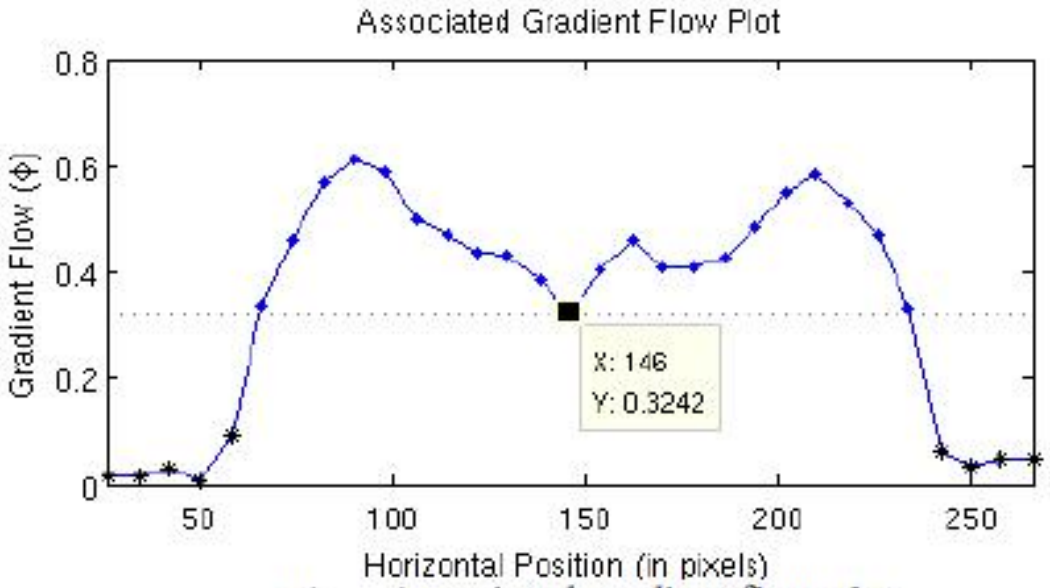
Juan-Bernardo Gómez-Mendoza

Abstract

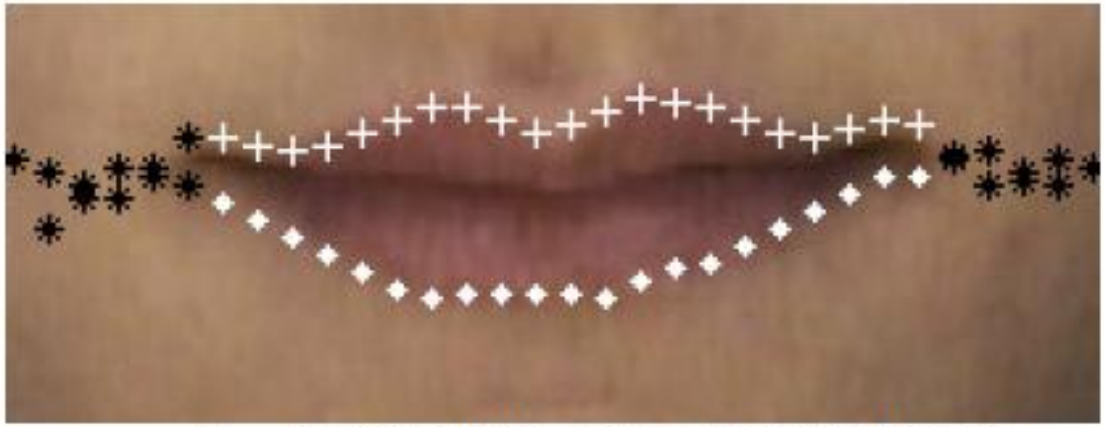
„In this paper a new method for outer lip contour approximation in facial images is presented.

*The method is inspired in the **Jumping Snakes** improving the later in terms of snake final position refinement. „*

Metoda wykorzystuje metodę tzw. Jumping snakes, dzięki czemu wyznacza charakterystyczne punkty i krzywe na konturach ust. Zmiana położenia poszczególnych punktów, ich obwiednia i późniejsza jej analiza umożliwia wyznaczenie odpowiedniego fonemu odpowiadającemu danemu kształtowi ust.



a) Associated gradient flow plot.



b) Contour plots showing untrimmed points.

Figure 1. Associated gradient flow use for point trimming.

Podsumowanie

- Kilka algorytmów, założenia podobne
- Widoczne jest duże zainteresowanie analizą ruchu ust w celu wspomagania rozpoznawania mowy
- W świecie, gdzie hałas zdaje się być nieodzowną jego częścią metody te ułatwią komunikację i przesył informacji

Bibliografia

- „Audiowizualna Baza Nagrań Mowy Polskiej”
  M.Igras, B.Ziółko, T.Jadczyk
- „Speaker Independent Visual-only Language Identyfication” J.L.Newman and S.J.Cox
- „Lip Localization and Viseme Classification for Visual Speech Recognition”
  S.Werda, W.Mahdi, A.B.Hamadou
- „Metody przetwarzania obrazów z wykorzystaniem metody OpenCV” S.Wojas
- „Automatic Outer Lip Contour Extraction in Facial Images”J-B.Gómez-Mendoza

Dziękuję za uwagę!