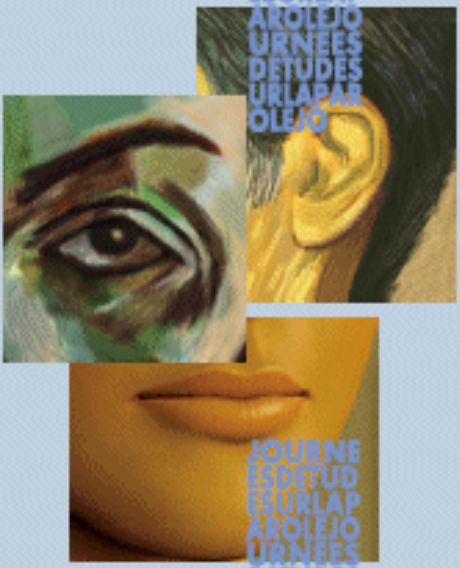
JEP 2006

XXVIes journées d'étude sur la parole

12-16 juin 2006
Dinard



Résumés des articles

AFCP

IRISA

ISCA

SESSION ORALE 1 : Introduction

Phonétique et Phonologie au siècle des Lumières

Christophe Rey.

The present paper proposes a presentation of various aspects of Nicolas Beauzée's theories on French sounds. Remained unexploited, Beauzée's theories sum up the best knowledge available in XVIII th century, before the progress brought by Comparatism and Dialectology, and found the description of the sounds of the language as a true field of study for grammatical science. We will also show that these same theories represent the first steps of a properly phonological reflexion.

À la poursuite de la trace du signal de parole

Bernard Teston.

At the beginning of the nineteenth century, the linguists, physiologists and acoustics experts had only one goal: to make the speech visible to be able to study its nature and its structure. Many scientists and inventors then will often launch out to the continuation of the speech signal with very varied but not very effective techniques, during meadows of a siécle. However, these sometimes curious devices will allow the researchers of this time to make fundamental discoveries on which our speech domain is founded.

Augmentation du taux de fausses acceptations par transformation inaudible de la voix des imposteurs

Jean-François Bonastre, Driss Matrouf, Corinne Fredouille.

This paper investigates the effect of a transfer function-based voice transformation on automatic speaker recognition system performance. We focus on increasing the impostor acceptance rate, by modifying the voice of an impostor in order to target a specific speaker. This paper is based on the following idea: in several applications and particularly in forensic situations, it is reasonable to think that some organizations have a knowledge on the speaker recognition method used and could impersonate a given, well known speaker. We also evaluate the effect of the voice transformation when the transformation is applied both on client and impostor trials.

EXPOSÉ INVITÉ

Le langage humain à la lumière de l'évolution

Jean-Louis Dessalles

L'étude de la structure du langage a longtemps été menée dans l'ignorance totale de la manière dont il s'est constitué. Or, le fait que la faculté de langage résulte d'une évolution biologique pose des contraintes fortes et soulève de nouvelles questions. Les modules du langage (phonologie digitale, intonation, marquage syntaxique, syntagmes, etc.) n'ont pas pu apparaître en même temps, et chaque étape devait constituer un stade fonctionnel et localement optimal. Dans ce cas, quelle est la fonction biologique des différents modules ? Pourquoi la taille de nos lexiques est-elle trois ordres de grandeur au-dessus de celle des répertoires animaux ? Pourquoi avons-nous au moins deux systèmes syntaxiques concurrents ? Pour quelle raison les organes du langage ont-ils évolué significativement chez l'émetteur (pharynx) et très peu chez le récepteur si c'est ce dernier qui profite de l'acte de communication ? Les progrès actuels sur l'évolution du langages nous permettent de proposer des réponses à ces questions.

SESSION ORALE 2 : Reconnaissance de la parole

Expériences de transcription automatique d'une langue rare

Thomas Pellegrini, Lori Lamel.

This work investigates automatic transcription of rare languages, where rare means that there are limited resources available in electronic form. In particular, some experiments on word decompounding for Amharic, as a means of compensating for the lack of textual data are described. A corpus-based decompounding algorithm has been applied to a 4.6M word corpus. Compounding in Amharic was found to result from the addition of prefixes and suffixes. Using seven frequent affixes reduces the out of vocabulary rate from 7.0% to 4.8% and total number of lexemes from 133k to 119k. Preliminary attempts at recombining the morphemes into words results in a slight decrease in word error rate relative to that obtained with a full word representation.

Reconnaissance automatique de la parole en langue somalienne

Abdillahi Nimaan, Pascal Nocera, Jean-François Bonastre.

Most African countries follow an oral tradition system to transmit their cultural, scientific and historic heritage through generations. This ancestral knowledge accumulated during centuries is today threatened of disappearing. Automatic transciption and indexing tools seem potential solution to preserve it. This paper presents the first results of automatic speech recognition (ASR) of Djibouti languages in order to index the Djibouti cultural heritage. This work is dedicated to process Somali language, which represents half of the targeted Djiboutian audio archives. We describe the principal characteristics of audio (10 hours) and textual (3M words) training corpora collected and the first ASR results of this language. Using the the specificities of the Somali language, (words are composed of a concatenation of sub-words called ``roots'' in this paper), we improve the obtained results. We will also discuss future ways of research like roots indexing of audio archives.

Ancres macrophonétiques pour la transcription automatique

Daniel Moraru, Guillaume Gravier.

Automatic speech recognition mainly rely on hidden Markov models (HMM) which makes little use of phonetic knowledge. As an alternative, landmark based recognizers rely mainly on precise phonetic knowledge and exploit distinctive features. We propose a theoretical framework to combine both approaches by introducing prior (phonetic) knowledge in a non stationary HMM decoder. As a case study, we investigate how broad phonetic landmarks can be used to improve a HMM decoder by focusing the best path search. We show that every broad phonetic class bring a small improvement, the best improvement being obtained with glides. Using all broad phonetic classes brings a significant improvement by reducing the error rate from 22% to 14% on a broadcast news transcription task. We also experimentally demonstrate that landmarks does not need to be detected with precise boundaries and can be used to fasten the beam search algorithm.

SESSION POSTER 1

Détection et correction automatique des déviations dans la réalisation de l'accent lexical anglais par des apprenants français

Guillaume Henry, Anne Bonneau, Vincent Colotte.

The work presented here is developed within a project devoted to the acquisition of English prosody by French learners, using speech technology modifications, and knowledge about L1 and L2 prosody. Our goal is to provide learners with relevant feedback in an automatic manner by comparing the prosodic cues of their realizations to that of a model. We focus on the English lexical accent and propose methods to correct automatically the learner’s realizations. We present our strategy and illustrate it through a concrete example.

Perception de la colère dans un corpus de français spontané par des apprenants portugais et tchèques

Sophie de Abreu, Catherine Mathon, Daniela Perekopska.

The aim of this paper is to show how prosody can provide a sufficient amount of information which allows recognizing the emotion of anger in a French spontaneous corpus for foreign learners of French. We present the results of a perception test carried on two groups of foreign learners of French, Portuguese and Czech. They had to listen to French sentences and evaluate the presence, or lack thereof, and the degree of anger in these sentences. We chose to use a real spontaneous corpus in order to keep the intonation of emotions in French intact. The semantic content was neutralised to put aside the information given by the content.

La production et la perception des voyelles orales françaises par les apprenants japonophones

Takeki Kamiyama.

In order to examine the production and perception of French oral vowels by native speakers of Japanese learning French as a foreign language, a series of experiments were conducted. First, 10 isolated oral vowels pronounced by 4 native speakers of French (2 male and 2 female) were identified by 5 Japanesespeaking learners. Second, the formant frequencies were measured for the vowels that were 1) read, and 2) repeated after native speakers' recordings, by 3 learners. The results suggest that it is difficult to perceive and produce in a native-like manner not only "new" vowels (front rounded series) but also a "similar" one (such as the French high back vowel /u/), as well as open-mid / close-mid oppositions.

Reconnaissance de la parole guidée par des transcriptions approchées

Benjamin Lecouteux, Georges Linarès, Pascal Nocera, Jean-François Bonastre.

In many cases, a descriptive transcript can be associated to speech signal : movies subtitles, scenario and theatre, summaries and radio broadcast, rearranged transcription for political debates. Transcripts correspond rarely to the exact word utterance. An actor plays differently each play and a speaker is sometimes far of the promptor. The goal of this work is whether to align the given transcript when it matches the speech signal or to fall back on an automatic speech recognition (ASR) adapted with the transcript language. There are multiple applications : to help deaf people following a play with closed caption aligned to the voice signal (with respect to performers variations), to watch a movie in another language using aligned closed caption, to transcript in real time debates or meetings.

Détection automatique d'opinions dans des corpus de messages oraux

Nathalie Camelin, Géraldine Damnati, Frédéric Béchet, Renato De Mori.

Telephone surveys are often used by Customer Services to evaluate their clients' satisfaction and to improve their services. Large amounts of data are collected to observe the evolution of customers' opinions. Within this context, the automatization of the process of these databases becomes a crucial issue. This paper addresses the automatic analysis of audio messages where customers are asked to give their opinion over several dimensions about a Customer Service. Interpretation methods that integrate automatically and manually acquired knowledge are proposed. A set of classifiers with several input features is introduced for each type of knowledge to add robustness in reducing the effect of limitation of machine learning algorithms. Manual and automatic learning procedures are proposed for conceiving strategies for using the classifiers. Experimental results, done on a database collected from a deployed Customer Service in real conditions with real customers, show the benefits of the proposed strategies.

Estimation rapide de modèles de Markov semi-continus discriminants

Georges Linares, Christophe Levy, Jean-Christophe Plaignol.

In this paper, we present a fast estimation rule for MMIE (Maximum Mutual Information Estimation) training of semi-continuous HMM (Hidden Markov Models). We first present the method proposed by Povey et al. (1999) for weigth re-estimation. Then, the weight updating rule is formulated in the specific framework of semi-continuous models. Finally, we propose an approximated updating function which requires very low computational ressources. The first experiment validates this method by comparing our fast MMI estimator (FMMIE) and original one. We observe that, on a task of digit recognition, FMMIE obtains similar results than whose obtained by using the full updating rule. Then, semi-continuous models are integrated in a Large Vocabulary Continuous Speech Recognition (LVCSR) system. We use the real-time engine which has been involved in the ESTER evaluation campaign. Results show that the proposed MMIE models outperform significantly the system based on semi-continuous models and MLE training, while reducting the model complexity.

Deux stratégies articulatoires pour la réalisation du contraste acoustique des sibilantes /s/ et /S/ en français

Martine Toda.

This paper reports two articulatory strategies used in the realization of /s/ - /S/ contrast in French from the observation of the MRI data of seven native speakers. These strategies are: 'tongue position adjustment' and 'tongue shape adjustment'. By examining the articulation of the subjects whose frication noise was deviant, it appeared that they already used all the possibilities for compensation within their articulatory strategy. A better normalization of their frication noise would have required complex gestures (e.g. tongue backing AND doming), which are presumably avoided by virtue of articulatory economy.

Étude acoustique et articulatoire de la parole Lombard : Effets globaux sur l'énoncé entier

Maëva Garnier, Lucie Bailly, Marion Dohen, Pauline Welby, Helène Loevenbruck.

This study aims at characterizing the articulatory modifications that occur in speech in noisy environments, and at examining them as compensatory strategies. Audio, EGG and video signals were recorded for a female native speaker of French. The corpus consisted of short sentences with a Subject-Verb-Object (SVO) structure. The sentences were recorded in three conditions : silence, 85dB white noise, 85dB "cocktail party" noise. Labial parameters were extracted from the video data. The analyses enabled us to examine the effect of the type of noise and to show that hyper-articulation concerns lip aperture and spreading rather than lip pinching. The analysis of the relationship between acoustic and articulatory parameters show that this speaker especially adapts to noise not only by talking louder or increasing vowel recognition cues but also by increasing spectral emergence.

Locus equation pour les consonnes /b/, /d/ et /x/ du vietnamien

Eric Castelli, Anne Hierholtz.

Locus equation measurements are one approach used to characterise vocal tract resonances during stop consonant production, the place of articulation of the consonant and the nature of the vowel-consonant transition. Taking up again previous literature studies, the aim of the present work is to applying locus equation measurements to Vietnamese language, for two stop voiced consonants /b/ and /d/, and for the specific velar voiced fricative consonant. Because Vietnamese is a tonal language, a specific corpus was built, in order to avoid as much as possible tonal co-articulation effects. We also take into account the two specific vowels that present dynamic characteristics. Comparisons with other languages are given at the end of this study.

Étude de la réduction non linéaire de la dimension du signal de parole en vue de modélisations discriminatives par SVM

José Anibal Arias, Régine André-Obrecht, Jérôme Farinas, Julien Pinquier.

In this article we study some results of the non-linear dimensionality reduction of speech vectors. Spectral clustering, Kernel PCA, Isomap, Laplacian eigenmaps and Locally Linear Embedding are related non-supervised methods that help to discover important caracteristics from data such as high-density regions or low-dimensional surfaces (manifolds). This reduction of dimension is a necessary step when we whant to model speech sequences with discriminative functions such as Support Vector Machines.

Estimation de la fréquence des formants basée sur une transformée en ondelettes complexes

Laurence Cnockaert, Jean Schoentgen, Francis Grenez.

The objective of this paper is to evaluate the performances of a formant estimation method in tracking variations due to the vocal tract movement during the production of sustained vowel. The formant frequency estimation is based on the instantaneous frequency obtained by means of a complex wavelet transform and is synchronised with the glottal cycle. Results for synthetic speech signals show that the precision of the formant frequency estimation is high. However, the estimated results are influenced by variations of the vocal frequency and variations of close formants. The method is illustrated for real speech.

Un détecteur d'activité vocale visuel pour résoudre le problème des permutations en séparation de source de parole dans un mélange convolutif

Bertrand Rivet, Christine Servière, Laurent Girin, Dinh-Tuan Pham, Christian Jutten.

Audio-visual speech source separation consists in mixing visual speech processing techniques (e.g. lip parameters tracking) with source separation methods to improve the extraction of a speech of interest from a mixture of acoustic signals. In this paper, we present a new approach that combines visual information with separation method based on the sparseness of speech: visual information is used as a voice activity detector (VAD) which is plugged on acoustic separation techniques. Results show the efficiency of the approach in the difficult case of realistic convolutive mixtures. Moreover, the overall process is quite simpler than previously proposed audiovisual separation schemes.

Étude de la structure formantique des voyelles produites par des locuteurs

Fabrice Hirsch, Véronique Ferbach-Hecker, Florence Fauvet, Béatrice Vaxelaire.

The aim of this study is to analyse the steady—state portion of the first two formants (F1) and (F2) in the production of [pVp] sequences, containing vowels [i, a, u] pronounced in two speech rates (normal and fast) by groups of untreated and treated stutterers, and nonstutterer controls. Comparing data between treated or untreated stutterers and controls, a reduction of vowel space is observed for stutterers in a normal speaking rate. When speech rate increases, no reduction of vowel space is noticeable for untreated stutterers, contrary to treated stutterers and controls.

Modélisation Statistique et Informations Pertinentes pour la Caractérisation

Gilles Pouchoulin, Corinne Fredouille, Jean-François Bonastre, Alain Ghio, Marion Azzarello, Giovanni Antoine.

This paper investigates the class of information relevant for the task of automatic classification of pathological voices. By using a GMM-based classification system (derived from the Automatic Speaker Recognition domain), the focus was made on three main classes of information : energetic, voiced, and phonetic information. Experiments made on a pathological corpus (dysphonia) have shown that phonetic information is particularly interesting in this context since it permits to refine the selection of the relevant information by looking at phonem- or phonem classlevel (e.g. nasal vowels).

Parler femme et parler homme en japonais actuel: Formes terminales et indices prosodiques

Yukihiro Nishinuma, Akiko Hayashi, Hiroko Yabe.

This work reports findings on the relationship between speaker sex and linguistic behavior of young Japanese in explanation-giving dialogues. The relationship between speaker sex and (a) the choice of utterance final forms; (b) the prosodic characteristics on these forms, has thus been examined. Data from 110 students of the Tokyo area revealed no statistically significant effect of the sex factor in the linguistic forms used. However utterance final syllables had a statistically significant effect both on intonation and on rhythm.

EXPOSÉ INVITÉ

Analyse de la violence verbale : Quelques principes méthodologiques

Claudine Moïse.

Si l'on peut la décrire à partir d'actes de langage repérables et analysables (insulte, mépris, dénigrement, menace, etc.), dans une perspective descriptive (Lagorgette, D. Ogier, C. Rosier, L.), voire d'analyse conversationnelle (Traverso, V., Kerbrat-Orecchioni, C., Vincent, D.), la violence verbale doit être aussi appréhendée dans sa globalité (et au-delà du fait linguistique). Les pratiques discursives et les interactions renseignent sur les pratiques sociales, c'est-à-dire que, par les jeux utilisés dans l'interaction, les locuteurs donnent sens à leurs actions, aux prises de pouvoir et aux positionnements sociaux (Gumperz, J. Martin-Jones, M., Gal S., Heller, M.). Les formes adoptées dans l'interaction violente peuvent donc être comprises comme des actes individuels, adhésion ou distanciation par rapport à l'interlocuteur et mais aussi comme des actes socialement inscrits, signes d'identification, d¹appartenance ou de résistance. Nous avons défini la violence verbale comme des " montées en tension interactionnelle" marquées par des " déclencheurs " spécifiques, processus qui s'inscrit dans des rapports de domination entre les locuteurs, des télescopages

de normes et de rituels, des constructions identitaires. Il s'agit alors de dire une sociolinguistique du sujet, en oeuvre dans les interactions violentes. Cette présentation pourrait rendre compte des différents aspects abordés dans nos analyses et dans notre compréhension de la violence verbale, analyse en actes de langage et en montées en tension dans une perspective intersubjective. De cette façon, la violence verbale se structure en étapes séquentielles où contextes d'énonciation et significances culturelles et relationnelles jouent un rôle essentiel. D'un point de vue méthodologique, nous rendrons compte aussi comment la construction de l'analyse est sujette aux conditions d'élaboration et à la diversité des corpus, et à la place de l'enquêteur en prise avec la violence verbale.

SESSION ORALE 3 : PROSODIE

Lecture silencieuse et oralisée des phrases relatives : le rôle de la prosodie

Christelle Dodane, Angèle Brunellière.

The purpose of this study was to determine whether prosody contribute to the integration of the syntactic level during the reading of relative sentences. A behavioural experiment was run on 20 French subjects. First, they had to read sentences silently presented visually (without prosodic markers such as commas). In an additional task using the same procedure, 10 subjects have to read sentences in a loud voice. The comparison between the two tasks reveals that words with the greater reading time in the first study correspond in the second study to a major prosodic phrase boundary. Results suggest that subjects have to restore prosodic contour in order to process syntactically relative sentences. The study of prosody in such a task provides a new methodology to access to the on-line syntactic processing.

La courbe de F0 des sonantes initiales de syllabe joue-t-elle un rôle prosodique ? Étude-pilote de données d'anglais britannique

Alexis Michaud, Barbara Kühnert.

Several recent publications raise the issue whether the F0 curve of syllable-initial sonorants can play a prosodic role. The experimental evidence adduced in the present pilot study consists of 15 C1VC2 words, where C1 = /p/, /b/ or /m/, V = /a:/, /i:/, /u:/, and C2 = /t/; these words were said twice inside a carrier sentence by 4 speakers of British English. Comparison of the F0 curves of the /m/-initial syllables with those of the obstruent-initial syllables suggests that only the part of the F0 curve which corresponds to the syllable rhyme is to be taken into account at the stage of the interpretation of the word's F0 curve.

Le focus prosodique n'est pas que déictique : le modèle VID (Valence-Intensité-Domaine)

Véronique Aubergé, Albert Rilliard.

This paper summarizes several perception experiments showing that the morphology of the prosodic focus conveys more information than the only deictic information: (1) the binary valence - yes/no focus – which is perceptively quite categorical (a magnet effect is clear on the basis of an identification and a discrimination experiment [1]), (2) the intensity information, used by the speaker to give his preference for one of two focused elements, (3) the information of the focus domain, that are some segmentation cues about the focused element (phonological unit or word unit), which are perceptively identified by listeners. The morphological cues revealing Valence-Intensity- Domain are observed in particular in morphing procedure making clear the thresholds of quitecategorical behaviors.

SESSION POSTER 2

Représentation acoustique compacte pour un système de reconnaissance de la parole embarquée

Christophe Lévy, Georges Linarès, Jean-François Bonastre.

Speech recognition applications are known to require a significant amount of resources (training data, memory, computing power). However, the targeted context of this work -mobile phone embedded speech recognition system- only authorizes few resources. In order to fit the resource constraints, an approach based on a HMM system using a GMM-based state-independent acoustic modeling is proposed in this paper. A transformation is computed and applied to the global GMM in order to obtain each of the HMM state-dependent probability density functions. The proposed approach is evaluated on a French digit

recognition task. Our method leads a Digit Error Rate (DER) of 2%, when the system respects the resource constraints. Compared to an HMM with comparable resource, our approach achieved a DER relative decrease more than 50%.

Mesures de confiance trame-synchrone

Joseph Razik, Odile Mella, Dominique Fohr, Jean-Paul Haton.

This paper presents some confidence measures for large vocabulary speech recognition which can be evaluated directly within the first steps of the recognition process. Having some clues to drive the recognition process may help to improve the accuracy of the provided sentence. Confidence measures may fit to this goal, so we propose some measures that can help the engine as soon as possible, without having to wait for the recognition process to be completed. Furthermore, our confidence measures are local and they are based on partial word graphes. Experiments on a French broadcast news corpus are also presented and show results close to the post calculated version of the measures.

Reconnaissance robuste de parole en environnement réel à l'aide d'un réseau de microphones à formation de voie adaptative basée sur un critère des N-best Vraisemblances Maximales

Luca Brayda, Christian Wellekens, Maurizio Omologo.

Distant-talking speech recognition in noisy environnements is generally tackled by using a microphone array and a related multi-channel processing. Based on that framework, this paper proposes an N-best extension of the Limabeam algorithm, that is an adaptive maximum likelihood beamformer. N-best hypothesized transcriptions are generated at a first recognition step and then optimized independently one to each other. As a result, the N-best list is re-ranked, which allows selection of the maximally likely transcription to clean speech models. Results on real data show improvements over both Delay and Sum Beamforming and Unsupervised Limabeam at low SNR and with moderate reverberation.

Les nasales du portugais et du français : une étude comparative sur les données EMMA

Solange Rossato, António Teixeira, Liliana Ferreira.

In this paper we present a first comparative study of velum height and movement in French and Portuguese based on EMMA data. Results show that the velum height reaches the highest position for oral consonants and decreases for oral vowels, nasal consonants and nasal vowels for both languages. Open vowels were found to be pronounced with velum height similar to the height used in nasal consonants production. Nasal vowels are produced with the lowest velum height in both languages but reveal different dynamic patterns.

Indices acoustiques de la coarticulation bidirectionnelle dans les séquences VCV en arabe

Mohamed Embarki.

This study assessed anticipatory and carry-over coarticulation effects in contemporary standard Arabic. VCV pairs with non pharyngealized dental-alveolar consonants and their pharyngealized cognates were used. Each consonant was inserted in symmetric vocalic contexts [C], [C] and [C]). F2 was measured at V1mid, V1offset, V2onset and V2mid. The results showed carry-over effects with non pharyngealized consonants and anticipatory coarticulation in pharyngealized context.

Équation de locus comme indice de distinction consonantique pharyngalisé vs non pharyngalisé en arabe

Mohamed Embarki, Christian Guilleminot, Mohamed Yeou.

Locus equations are linear regression functions derived by relating F2 onsets of different vowels to their corresponding steady states. This paper purports to investigate if locus equations can be strong phonetic descriptors of the consonantal contrast between pharyngealized and non-pharyngealized consonants in Arabic. Eight male Arabic speakers from eight different Arabic countries produced 24 #CV# tokens, where C was either non-pharyngealized or pharyngealized. Each consonant was followed by one of the three vowels [i], [u] and [a].

Paramétrisation de la parole basée sur une modélisation des filtres cochléaires : application au RAP

Zied Hajaiej, Kaïs Ouni, Noureddine Ellouze.

Signal processing front end for extracting the feature set is an important stage in any speech recognition system. The optimum feature set is still not yet decided. There are many types of features, which are derived differently and have good impact on the recognition rate. This paper presents one more successful technique to extract the feature set from a speech signal, which can be used in speech recognition systems. Our technique based on the human auditory system characteristics and relies on the gammachirp filterbank to emulate asymmetric frequency response and level dependent frequency response. For evaluation a comparative study was operated with standard MFCC and PLP.

Vers un inventaire ordonné des configurations manuelles de la LSF

Leïla Boutora.

This article deals with French Sign Language (FSL), and particularly with the question of the definition of its minimal units (in realisation and perception). The general aim of this work is to know if we can make strict equivalence between phonems, the minimal units of realisation of vocalic languages (VL), and the minimal units of sign languages (SL). We make a focus on the status of handshapes in FSL in the lexicon and on the problem we are faced with the definition of their inventory in « phonetic » terms.

Identification perceptive d'accents étrangers en français

Bianca Vieru-Dimulescu, Philippe Boula de Mareüil.

A perceptual experiment was designed to determine to what extent naïve French listeners are able to identify foreign accents in French: Arabic, English, German, Italian, Portuguese and Spanish. They succeed in recognising the speaker's mother tongue in more than 50% of cases (while rating their degree of accentedness as average). They perform best with Arabic speakers and worst with Portuguese speakers. The Spanish and Italian accents on the one hand and the English and German accents on the other hand are the most mistaken ones. Phonetic analyses were conducted; clustering and scaling techniques were applied to the results, and were related to the listeners' reactions that were recorded during the test. Emphasis was laid on differences in the vowel realisation (especially concerning the phoneme /y/).

À propos du trait ATR des voyelles nasales du twi

Kofi Adu Manyah.

This paper investigates acoustic properties of the Twi nasal vowel counterparts of the oral vowels /i/ vs /I/ and /u/ vs /U/ investigated in a previous study of ATR vowel harmony. Acoustic measurements are carried out to investigate for differences between vowel quality in the 2 groups. The evidence from our acoustic data, confirming results obtained for the oral vowels, is the tendency for advanced vowels [+ATR] to have lower F1 values, higher F2 and F3 values than the unadvanced vowels.

Variation, coup de glotte et glottalisation en persan

Shahrbano, Suzanne Assadi.

Glottalization phenomena are produced with much variation across individual speakers. Data analysed here is consisted of 228 isolated words, 3 short texts and 2 dialogues (15 minutes), read by two native speakers of Persian. The paper describes the inter- and intra-speaker variability in the realization of glottalization, the effect of context and pitch accent.

Influence de la distribution et des caractéristiques acoustiques sur la perception des bilingues et des monolingues. Cas du /r/ chez les guadeloupéens et chez les français

Johanne Confiac-Akpossan.

First language shapes perception and production (Troubetzkoy, 1939; Kuhl et al., 1992). This paper compares Guadeloupean with French listeners. In Guadeloupe, creole and french coexist. The bilingual Guadeloupeans are exposed to two different phonological systems where consonant /r/ has different distribution and acoustic characteristics. Perceptual, phonological and statistical analyses tend to show an influence of these both parameters on speech perception.

Vous avez dit proéminence ?

Michel Morel, Anne Lacheret-Dujour, Chantal Lyche, François Poiré.

Analysing prosody requires the correct identification of prominence peaks. This paper examines the results of an experiment where 7 linguists submitted to a prominence identification task largely failed to agree. We show that prominence detection is proportionate to F0 variation, but not to length, that there exists considerable variation between judges, the best of whom barely attains a 50 % score of correct answers. We conclude that coding prosody in a large corpus will require the use of dedicated software to supplement the work done by individual coders.

Intonation des phrases interrogatives et affirmatives en langue vietnamienne

Minh-Quang Vu, Do-Dat Tran, Eric Castelli.

Interrogative and affirmative sentences in Vietnamese language, which have same tones and the same number of syllables, are recorded in order to analyse their intonation shape (F0 evolution) avoiding effect of lexical tones and of co-articulation. Comparisons permit us to characterise differences between question and non-question at sentence prosody level. Then, our work is completed by a perception study; its main goal is to check if sentence nature characteristics are included in the sentence prosody, allowing the auditor to classify questions and non-questions, despite the complex form of this prosody in tonal languages. Results show that information on sentence nature are present at the end of its last demi-syllable and that about 70% of sentences are good classified.

SESSION ORALE 4 : Reconnaissance du locuteur et de la langue

Identification automatique des parlers arabes par la prosodie

Jean-Luc Rouas, Mélissa Barkat-Defradas, François Pellegrino, Rym Hamdi-Sultan.

This paper presents a study of automatic identification of Arabic dialectal areas based on a prosodic automatic modelling. Inspired from Fujisaki's works, this modelling dissociates long term prosodic variations from short term micro-variations and exploits n-multigrams models. Experiments, achieved on semi-spontaneous recordings from 40 speakers, show that the system reaches 98% of correct identification of the three dialectal areas - Maghreb, Middle-East, and an intermediate area (Tunisia-Egypt) - with test excerpts of 7.6 seconds in average.

Identification automatique des langues : combinaison d'approches phonotactiques à base de treillis de phones et de syllabes

Dong Zhu, Martine Adda-Decker.

This paper investigates the use of phone and syllable lattices to automatic language identification (LID) based on multilingual phone sets (73, 50 and 35 phones). We focus on efficient combinations of both phonotactic approach and syllabotactic approaches. The LID structure used to achieve the best performance within this framework is similar to PPRLM (parallel phone recognition followed by language dependent modeling): several acoustic recognizers based on either multilingual phone or syllabe inventories, followed by languagespecific n-gram language models. A seven language broadcast news corpus is used for the development and the test of the LID systems. Our experiments show that the use of the lattice information significantly improves results over all system configurations and all test durations. Multiple system combinations further achieves improvements.

Application des machines à vecteurs support mono-classe à l'indexation en locuteurs de documents audio

Belkacem Fergani, Manuel Davy, Amrane Houacine.

This paper addresses a new approach based on the kernel change detection algorithm introduced recently by Desobry et al. This new algorithm is applied to the speaker change detection and clustering tasks, which are the key issues in any audio indexing process. We show the efficiency of the method through several experiments using RT'03S NIST data. We discuss also the parameters tuning and compare the results to the well known GLR-BIC algorithm.

Indexation en locuteur : utilisation d'informations lexicales

Julie Mauclair, Sylvain Meignier, Yannick Esteve.

The automatic speaker indexing consists in splitting the signal into homogeneous segments and clustering them by speakers. However the speaker segments are specified with anonymous labels. This paper propose to identify those speakers by extracting their full names pronounced in the show. With a semantic classification tree, the full names detected in the segment transcription are associated to this segment or to one of its neighbors. Then, a merging method associates a full name to a speaker cluster instead of the anonymous label. The experiments are carried out over French broadcast news from the ESTER 2005 evaluation campaign. About 70% show duration is correctly processed for the evaluation corpus.

Une nouvelle approche fondée sur les ondelettes pour la discrimination parole/musique

Emmanuel Didiot, Irina Illina, Odile Mella, Dominique Fohr, Jean-Paul Haton.

The problem of Speech/Music discrimination is a challenging research problem which significantly impacts Automatic Speech Recognition (ASR) performance. This paper proposes new features for the Speech/Music discrimination task. We use a decomposition of the audio signal based on wavelets which allows a good analysis of non stationary signals like speech or music. We compute different energy types in each frequency band obtained from wavelet decomposition. We use two Class/Non-Class classifiers : one for speech/non speech, one for music/non music. On a broadcast corpus, using the proposed wavelet approach, we obtained a significant improvement (35%) compared to MFCC parameters.

Représentation paramétrique des relations temporelles appliquée à l'analyse de données audio pour la mise en évidence de zones de parole conversationnelle

Zein Al Abidin Ibrahim, Isabelle Ferrané, Philippe Joly.

The general aim of our work is the automatic analysis of audiovisual document to caraterize their structure by studying the temporal relations between the events occurring in it. For this purpose, we have proposed a parametric representation of temporal relations. From this representation, a TRM (Temporal Relation Matrix) can be computed and analyzed to identify relevant relation class. In this paper, we applied our method on audio data, mainly on speaker and applause segmentations from a TV game program. Our purpose is to analyze these basic audio events, to see if the observations automatically highlighted could reveal information of a higher level like speaker exchanges or conversation, which may be relevant in a structuring or indexing process.

SESSION POSTERS 3

Généralisation du noyau GLDS pour la vérification du locuteur par SVM

Jérôme Louradour, Khalid Daoudi, Francis Bach.

The Generalized Linear discriminant Sequence (GLDS) kernel provides good performance in SVM speaker verification in NIST SRE (Speaker Recognition Evaluation) evaluations. It is based on an explicit mapping of each sequence to a single vector in a feature space using polynomial expansions. Because of

practical limitations, these expansions have to be of degree less or equal to 3. In this paper, we generalize the GLDS kernel to allow not only any polynomial degree but also any expansion (possibly infinite dimensional) that defines a Mercer kernel (such as the RBF kernel). We conceive a new kernel, and makes it tractable using a method of data reduction adapted to kernel methods : the Incomplete Cholesky Decomposition (ICD). We present experiments on NIST SRE database, that show good perspective for our new approach

Représentation du locuteur par modèles d'ancrage pour l'indexation de documents audio

Mikaël Collet, Delphine Charlet, Frédéric Bimbot.

This paper presents a speaker indexing system of audio document entirely based on the anchor models approach. Evaluation is done on the audio database of the ESTER evaluation campaign for the rich transcription of French broadcast news. Results show that speaker indexing performances are improved when a speaker clustering process is performed and that a weighted measure of similarity, used in the speaker tracking process, can overcome some errors of the clustering process. The use of anchor models is

particulary suitable for speaker indexing because the computational burden to search a speaker in an audio document is very low and performances are equivalent to those of a speaker indexing system using the classical speaker representation in the acoustic space (Gaussian model for speaker segmentation and clustering, Gaussian mixture model for speaker tracking).

Application d'un algorithme génétique à la synthèse d'un prétraitement non linéaire pour la segmentation et le regroupement du locuteur

Christophe Charbuillet, Bruno Gas, Mohamed Chetouani, Jean-Luc Zarader.

Speech feature extraction plays a major role in a speaker recognition system. B. Gas & al. showed in [1] that a non linear filtering of speech can improve the feature extractor's ability. In this article we propose to use genetic algorithms to design a non-linear pre-processing of speech adapted to the speaker diarization task. The pre-processing system we present is based on artificial recurrent neural networks (ARNN). We used a genetic algorithm to find both the structure and the weights of the network. Experiments are carried out using a state-of-the-art speaker diarization system. Results showed that the proposed method give significant improvements, reducing the diarization error rate from 17.38 % to 15.77 %.

Influence de la corrélation entre le pitch et les paramètres acoustiques en reconnaissance de la parole

Gwenael Cloarec, Denis Jouvet, Jean Monne.

In this paper we compare the role played by the pitch frequency on speaker independent speech recognition performances for two tasks: an isolated word recognition task and a continuous speech recognition task. While introducing pitch and/or voicing directly into the acoustic vector leads to significant improvements on the isolated word recognition task, this method does not bring any improvement on the continuous speech recognition task. On the contrary, modelling the pitch frequency independently of the acoustic parameters leads to small but similar improvements on the two tasks. Those results could be explained by the fact that the improvement brought when introducing pitch and voicing directly into the acoustic vector is related to the correlation between the pitch frequency and the acoustic parameters. This correlation is much less important in the case of the continuous speech recognition task in which prosody can lead to very different pitch values depending on the prosodic context.

Transformation linéaire discriminante pour l'apprentissage des HMM à analyse factorielle.

Fabrice Lefèvre, Jean-Luc Gauvain.

Factor analysis has been recently used to model the covariance of the feature vector in speech recognition systems. Maximum likelihood estimation of the parameters of factor analyzed HMMs (FAHMMs) is usually done via the EM algorithm. The initial estimates of the model parameters is then a key issue for their correct training. In this paper we report on experiments showing some evidence that the use of a discriminative criterion to initialize the FAHMM maximum likelihood parameter estimation can be

effective. The proposed approach relies on the estimation of a discriminant linear transformation to provide initial values for the factor loading matrices. Solutions for the appropriate initializations of the other model parameters are presented as well. Speech recognition experiments were carried out on the Wall Street Journal LVCSR task with a 65k vocabulary. Contrastive results are reported with various model sizes using discriminant and non discriminant initialization.

Proposition d'une nouvelle méthodologie pour la sélection automatique du vocabulaire d'un système de reconnaissance automatique de la parole

Brigitte Bigi.

The vocabulary of an Automatic Speech Recognition (ASR) system is a significant factor indetermining its performance. The goal of vocabulary selection is to construct a vocabulary with exactly those words that are the most likely to appear in the test data. This paper proposes a new measure to evaluate the quality of a vocabulary regarding a domain-specific ASR application. This Q_alpha-measure is based on the trade off between the target lexical coverage and vocabulary size. Experiments were carried out on French Broadcast News Transcriptions using the Q_alpha-measure compared to the state-of-the-art method. Results of these two methods favor systematically the proposed methodology.

Théorie de la syllabe et durées vocaliques : vers une interprétation unifiée du rôle de la structure syllabique et de la nature des segments.

Olivier Crouzet, Angoujard Jean-Pierre.

It is generally agreed that vowel duration may be influenced by both phonetic context and syllable structure. Though it may seem reasonable to call for two different variables in this respect, we show that the rythm and substance approach to syllable structure may offer a common framework for unifying these two sources of variation into a single theoretical account. The results of a speech production experiment involving french speakers are described which confirm that this syllabic approach accounts for some of the predicted deviations in vowel duration when syllabic and contextual effects are involved. Though further studies should be conducted, this framework seems particularly promising for the understanding of the relationship between articulatory, phonological and rythmic influences on speech production mechanisms.

Effets aérodynamiques du mouvement du velum : le cas des voyelles nasales du français

Amelot Angélique, Michaud Alexis.

Hitherto unpublished data on oral airflow, nasal airflow and velum movement during logatoms read by two French speakers allow for the investigation of the relationships between these three phenomena. There is no straightforward relationship between velar movements and nasal airflow, the latter depending on the relative impedance of both tracts, reflected in the ratio of nasal airflow to oral airflow. The structure of the 168 logatoms is C1V1C1VtC1V1, where C1 = /t/, /d/, /l/, /n/, /s/ or /z/, V1 = /a/, /i/, /u/ or /y/, and Vt = /E~/, /A~/, /O~/, /a/, i/, /u/ or /y/, allowing for a characterisation of the effect of these consonantal and vocalic contexts on airflow and velum movement. In particular, a hypothesis is put forward concerning the frequent dip below zero of nasal airflow after stop consonants, and its effect on oral airflow.

Sensibilité au débit et marquage accentuel des phonèmes en français

Valérie Pasdeloup, Robert Espesser, Malika Faraj.

The aim of this work is to determine the way the prosodic scene reorganises itself according to speech rate variations in French. We present the temporal structure study of a one thousand word speech corpus. The corpus was produced at three different rates (normal, fast and slow) by one speaker with two repetitions. The goal is to study the relationship between speech rate sensitivity and accentual markedness of phoneme. Results put in light that phoneme does not behave the same way if stressed or not and if consonantic or vocalic. Unstressed phonemes are less rate sensitive than stressed ones. Vowels are more rate sensitive than consonants, especially when stressed. Nevertheless, consonants are rate sensitive in such proportions when stressed that it is not possible to say, as usually said, that it is the vowel which carries stress.

Différentiation des mots de fonction et des mots de contenu par la prosodie : analyse d'un corpus trilingue de langage adressé à l'enfant et à l'adulte.

Christelle Dodane, Jean-Marc Blanc, Peter Ford Dominey.

This research investigated the role of salient prosodic cues in the first approximant assignment in function and content words by infants. In order to discover these cues and to see if they could vary cross-linguistically, infantdirected speech was compared to adult-directed speech in three different languages, French, English and Japanese. The same story was successively read by 15 mothers to their infant and to an adult (5 mothers for each language). The acoustic analyses reveal that among 33 different prosodic cues, non final syllable duration, Fo peaks and amplitude peaks relevant in the three languages to allow categorization in function and content items (Fo peaks, amplitude peaks and non-final syllable duration), but they have a different relative weight across languages because of the specific prosodic organization of these languages.

Comment les attitudes prosodiques sont parfois de « faux-amis » : les affects sociaux du japonais vs. français

Takaaki Shochi, Véronique Aubergé, Albert Rilliard.

The attitudes of the speaker during a verbal interaction are affects linked to the speaker's intentions, and are built by the language and the culture. Attitudes are the main part of the affects expressed during everyday interactions. This paper describes several experiments underlying that some attitudes belong both to Japanese and French languages, and are implemented in perceptively similar prosodies, but that some Japanese attitudes don't exist and/or are wrongly decoded by French listeners. Results are presented for the 12 attitudes and three levels of language learning (naive, beginner, intermediary). It has to be noted that French listeners, naive in Japanese, can very well recognize admiration, authority and irritation; that they don't discriminate Japanese question from declaration before the intermediary level, and that the extreme Japanese politeness is interpreted as impoliteness by French listeners, even when they can speak a good level of Japanese.

Expressions hors des tours de parole : éthogrammes du « feeling of thinking »

Fanny Loyau, Véronique Aubergé, Anne Vanpé.

During our collect of an expressive corpus, a large quantity of non verbal information has been registered too: top body and face movements, and voice events. We are particularly interested by only these actions which happen outside the talk turn, when the subject thinks, and feels about what he thinks. We want to know if these events are real indices of signals about the mental states or the affective states of the subjects. For that, a typical ethogram methodology has been applied to label these non speech parts into primitive icons of top body movements, face movements and voice events, in order not to take any decision about the interpretation of what could be expressed by these events, but to classify variant movements into minimal icons.

Acquisition de la liaison chez l'enfant francophone : formes lexicales de Mots2

Céline Dugua, Damien Chabanal.

The aim of this research is to establish which lexical forms in Word2 position are available to the child in liaison processes. It is known that the child makes many errors at the beginning of the acquisition of liaison and elision, of the sort "la noreille, le zours, petit néléphant, un zéléphant…". These variants argue in favour of the encoding of the liaison consonant at the beginning of Word2 in the early stages of the acquisition of liaison. To test this hypothesis, we present the results of two studies (one transversal, the other longitudinal) of French-speaking children between 2;6 and 6;3. The data collected allow us to conclude that the children have, at the outset, in their mental lexicon, Word2 forms with both an initial consonant and an initial vowel. These observations suggest that the child posseses several exemplars of the same Word2.

Changements intonatifs dans la parole Lombard : au-delà de l'étendue de F0

Pauline Welby.

Earlier studies of speech in noise (Lombard speech) have generally reported an increase in fundamental frequency (F0). This study examined other potential intonational differences. Seven French speakers read a corpus of short paragraphs, in quiet and in 80 dB white noise. Four speakers increased F0 range across target accentual phrases in noise. Six upscaled individual tones, although there was great inter-speaker variability. Noise did not influence intonation pattern type; in particular, there was no tendency to produce more "early rises" in noise, even though these rises are cues to word segmentation. Producing an early rise (thus a LHLH or LHH pattern) may not add to the salience of the commonly produced LH pattern. There were no differences in tonal alignment, in contrast to earlier findings. This null result may be due to paradigm differences between the two experiments.

Le Paradigme ascendent de l'FO dans les fonctions préindicatives adverbiales en portugais brésilien

Cirineu Stein.

This paper presents the results of two perception tests, which aimed to identify, in Brazilian Portuguese, if a native user of the language, not trained in phonetics, is able to recognize a pre-indicative adverbial value, and, with the help of vocal synthesis techniques, which specifications of the prosodic components are responsible for the establishment of that value. Although the whole of the research focuses on nine

adverbial semantic fields, only three of them will be discussed here. These three possible pre-indicative patterns of cause, consequence and finality show an ascending melodic curve in the final stressed vowel. Due to the similarity among the melodic curve contours in these three pre-indicative patterns, it is possible that a careless listener perceives them as ambiguous.

EXPOSÉ INVITÉ

Towards Domain Unlimited Speech Translation

Tanja Schultz

This paper describes our ongoing work in domain unlim-ited speech translation. We describe how we developed a lecture translation system by moving from speech transla-tion of European Parliament Plenary Sessions and seminar talks to the open domain of lectures. We started with ourspeech recognition (ASR) and statistical machine translation (SMT) 2006 evaluation systems developed withinthe framework of TC-Star (Technology and Corpora for Speech to Speech Translation) and CHIL (Computers inthe Human Interaction Loop). The paper presents the speech translation performance of these systems on lec-tures and gives an overview of our final real-time lecture translation system.

SESSION ORALE 5 : Compréhension automatique

Mesure de confiance de relation sémantique dans le cadre d'un modèle de langage sémantique

Catherine Kobus, Géraldine Damnati, Lionel Delphin-Poulat

This article proposes a new confidence measure estimated for concept hypotheses given by a semantic language model. This confidence measure is based upon the ontology and the semantic relations linking concepts of a dialog application. It aims at measuring how high a concept hypothesis is related to the other hypotheses of an utterance. The semantic relation confidence measure is evaluated alone and in combination with a classical acoustic confidence measure. It is shown that the two confidence measures are complementary and yield good performance in terms of cross entropy relative reduction.

Décodage conceptuel à partir de graphes de mots sur le corpus de dialogue Homme-Machine MEDIA

Christophe Servan, Christian Raymond, Frédéric Béchet, Pascal Nocéra.

Within the framework of the French evaluation program MEDIA on spoken dialogue systems, this paper presents the methods proposed at the LIA for the robust extraction of basic conceptual constituents (or concepts) from an audio message. The conceptual decoding model proposed follows a stochastic paradigm and is directly integrated into the Automatic Speech Recognition (ASR) process. This approach allows us to keep the probabilistic search space on sequences of words produced by the ASR module and to project it to a probabilistic search space of sequences of concepts. The experiments carried on on the MEDIA corpus show that the performance reached by our approach is better than the trdiational sequential approach that looks first for the best sequence of words before looking for the best sequence of concepts.

Un modèle stochastique de compréhension de la parole à 2+1 niveaux

Hélène Bonneau-Maynard, Fabrice Lefèvre.

In this paper an extension is presented for the 2-level stochastic speech understanding model, previously introduced in the context of the Arise corpus. In the new model, an additional stochastic level is in charge of the attribute value normalization. Due to data sparseness, the full 3-level model is not applicable straightforwardly and a variant is introduced where the conceptual decoding and value normalization phases are decoupled. The proposed approach is evaluated on the French Evalda-Media task (hotel booking and tourist information). This recent corpus has the advantage to be semantically annotated with conceptual segments, which allows for a direct training of the 2-level model. We also present some further model improvements such as the modality propagation or the 2-step hierarchical recomposition. On the whole, the various proposed techniques reduce the understanding error rate from 37.6% to 28.8% on the development set (24% relative improvement). This model has been engaged in the 2005 Media evaluation campaign where it achieved the best results among the 5 participants with an error rate of 29%.

SESSION POSTER 4

Évaluation de systèmes de génération de mouvements faciaux

Oxana Govokhina, Gérard Bailly, Gaspard Breton, Paul Bagshaw.

This paper presents the implementation and evaluation of different movement generation techniques for speech-related facial movements. State-of-the-art systems are implemented. A novel system that combines HMM-driven pre-selection of diphones with a standard concatenation system is also implemented. The trajectory formation systems are parameterised using the same training material. The groundtruth data consists of facial motion and acoustic signals of one female speaker uttering 238 sentences. Both objective and subjective evaluation of the systems is reported. The objective evaluation observes the linear correlation coefficient between original and predicted movements. It is complemented by an audiovisual preference test where ground-truth and predicted movements drive a 3D virtual clone of the original speaker.

Contraintes globales pour la sélection des unités en synthèse vocale

Adrian Popescu, Cédric Boidin, Didier Cadic.

This work proposes an alternative unit selection method for corpus-based voice synthesis. It introduces the need of long term constraints in the cost function, which cannot be handled by the traditional Viterbi algorithm. Therefore another optimization algorithm, the simulated annealing, has been chosen for our experiments. It has been evaluated on a cost function encouraging long term F0 continuity. Although the results of our experiments do not show real improvement of the overall quality, they involve further research on this relevant issue.

Une synthèse vocale destinée aux déficients visuels

Hélène Collavizza, Jean-Paul Stromboni.

This paper presents an experiment in developping and testing a text to speech system which is needed for improving some applications dedicated to visually impaired users. When carrying out such applications, it appears that an interface able to speak, i.e. read a text, is mandatory. In order to allow an easy release, this text to speech system should be portable, licence free, using freewares and free solutions. To fulfill all these needs, a solution was chosen and developped. On the one hand, several experiments were conducted with different such applications. On the other hand, the obtained text to speech system was made available on the Web for evaluation purpose. These two information sources allow to gather a set of advices and knowledge on such tools, and make possible and easier to develop next versions.

Peut-on utiliser les étiqueteurs morphosyntaxiques pour améliorer la transcription automatique ?

Stéphane Huet, Guillaume Gravier, Pascale Sébillot.

The aim of the paper is to study the interest of part-of-speech (POS) tagging to improve speech recognition. We first evaluate the part of misrecognized words that can be corrected using POS information; an analysis of a short extract proves that a decrease of the word error rate by 1.1 point can be expected. We also demonstrate quantitatively that traditional POS taggers are reliable when applied to spoken corpus, including automatic transcriptions. This new result enables us to effectively use POS tag knowledge to improve, in a postprocessing stage, the quality of transcriptions, especially correcting agreement errors; these first preliminary results however still have to be bettered to lower the overall word error rate.

Algorithme de recherche d'un rang de prédiction. Application à l'évaluation de modèles de langage

Pierre Alain, Olivier Boeffard.

Within a predictive framework for language model evaluation, Shannon uses a rank distribution in order to bound the entropy of printed English. Taking into account of higher dimensions (prediction of symbols in a raw) and predicting a k-word sequence given a N-word vocabulary is a NP-hard computational task. To achieve this goal, we propose some acceptable and effective search heuristics for an A* algorithm.

Étude comparative de modélisation de langage par bigrams et par multigrams pour la reconnaissance de parole

Yassine Mami, Frédéric Bimbot.

The use of stochastic ngram models has a long and successful history in the research community; nowadays ngrams are becoming quite common in commercial systems, as the market demands more robust and flexible solutions. This approach is particularly interesting for its effectiveness and its robustness, but limited to modeling only local linguistic structures. To overcome this limitation, we propose the use of models with variable length. In this paper we present the multigram language models and we integrate them in a speech recognition system. The experiments are carried out on a France Telecom's dialogue application for stock exchange.

La prosodie des mots grammaticaux : le cas des deux déterminants "du" et "deux"

Takeki Kamiyama.

Does explicit knowledge of prosody help L2 learners to identify the two determiners "du" and "deux" in French? An analysis of 162 sentences read by 3 French native speakers show the expected tendency of F0 and duration ("deux" being longer and higher than the function word "du"). Then, 3 sets of 8 synthesised stimuli were generated using Mbrola, with expected and unexpected f0 and duration patterns. A perception experiment with 16 French native speakers suggests that they tend to be biased by the unexpected prosody (duration, in particular) when they listen to the sentences with white noise. In another experiment, three groups of Japanese-speaking learners were asked to identify the two words in 48 sentences read by a native speaker. The preliminary results suggest that teaching explicit knowledge of prosody might facilitate the acquisition.

Aspects phonologique et dynamique de la distinctivité au sein des systèmes vocaliques : une étude inter-langue

Christine Meunier, Robert Espesser, Cheryl Frenck-Mestre.

This study is a cross-linguistic investigation of qualitative and quantitative variations due to 1/ the structure of vocalic system, 2/ the amount of context within speech message. We hypothesize that phonetic distinctivity of vowels in a language is relative to 1/ the properties of the phonological system, 2/ the amount of informational context. Three languages (Spanish, French and English) were analyzed in three different types of speech (isolated vowels, within words and within texts). Results show 1/ centralization in the three vocalic systems relative to the amount of context, 2/ an increase of vowel dispersion also due to an increase of context information.

Natures de schwa en gallo

Jean-Pierre Angoujard.

In this paper we offer a first declarative analysis of Gallo schwa (Gallo is a Romance language spoken in eastern Britanny). Whereas the behaviour of schwa in French can be derived from the properties of a unique object, the various achievement of schwa in Gallo must be associated dictinct objects: a default vowel (as in French), a lexical vowel, and an optional vocoid preceding a syllabic consonant. We will show that the major properties of Gallos schwa(s) can be described through only two constraints, of which one is lexical and the other is rhythmic.

Dimensions acoustiques de la parole expressive : poids relatifs des paramètres resynthétisés par Praat vs. LF-ARX

Nicolas Audibert, Damien Vincent, Véronique Aubergé, Albert Rilliard, Olivier Rosec.

The emotional prosody is multi-dimensional. A debated question is whether some parameters are more specialized to convey some emotion dimensions. Selected stimuli carrying acted expressions of anxiety, disappointment, disgust, disquiet, joy, resignation, satisfaction and sadness on monosyllabic words were used to synthesize artefactual stimuli by projecting separately prosodic parameters on neutral expressions, with Praat and an LF-ARX algorithm. Perceptive evaluation of stimuli and comparison of results (1) indicate that F0 contours bring more information on positive expressions while voice quality and duration convey more information on negative expressions, and intensity alone is not informative enough (2) diagnoses minor artifacts of both synthesis methods which consequences may have interesting implications in expressive speech synthesis (3) validates the efficiency of the LF-ARX algorithm (4) measures the relative weights of each of the LF-ARX voice quality parameters.

Vers un système multilinéaire de transcription des variations intonatives

Berchtje Post, Elisabeth Delais-Roussarie.

In the paper, we will present a transcription system for Intonational Variation (IVTS), derived from IViE. The prosodic features are transcribed on i) the rhythmic tier ; ii) the local phonetic tier ; iii) the global phonetic tier ; and iv) the phonological tier. Each tier offers a range of labels which share a general architecture, but language-specific parameters determine which subset of labels a transcriber can choose from for the transcription of a particular language variety. In this paper, we will argue that the multi-linear architecture of IViE-based systems offers transparency, flexibility and standardization, three key advantages in qualitative and quantitative studies of intonational variation across languages and language varieties.

Relations entre le bruit entachant les paramètres de contrôle des modèles non linéaires et le bruit mesuré en sortie

Michel Pitermann.

Il arrive fréquemment que pour générer des simulations à l'aide d'un modèle non linéaire, des mesures de grandeurs physiques du monde réel soient utilisées comme valeurs des paramètres de contrôle du modèle. Dans ce cas, le bruit mesuré à la sortie du modèle contient au moins deux composantes : (i) un bruit d'origine chaotique intrinsèque au modèle ; (ii) un bruit provenant du bruit de mesure des grandeurs physiques extrinsèques au modèle. Une méthode pour quantifier la composante de bruit d'origine chaotique a été proposé dans [2]. Le présent article complète la méthode en proposant une technique destinée à quantifier la deuxième composante de bruit provenant du bruit entachant les paramètres de contrôle du modèle. Cette technique est illustrée par l'analyse d'un modèle biomécanique de visage. Les résultats montrent que malgré sa simplicité la méthode permet d'estimer correctement le bruit de sortie en fonction du bruit présent dans les paramètres de contrôle du modèle.

Modélisation physique des cordes vocales : comment tester la validité des modèles ?

Nicolas Ruty, Annemie Van Hirtum, Xavier Pelorson.

An experimental set-up and human vocal folds replica able to produce self sustained oscillations is presented. The aim of the set-up is to assess the relevance and the accuracy of theoretical vocal folds models. The applied reduced mechanical models are a variation of the classical two-mass model. The airflow is described as a laminar flow with flow separation. The influence of a downstream resonator is taken into account. The oscillation pressure threshold and fundamental frequency are predicted by applying a linear stability analysis to the mechanical models. The measured frequency response of the mechanical replica together with the initial (rest) area allows to determine the model parameters (spring stiffness, damping, geometry, masses). Validation of theoretical model predictions to experimental data shows the relevance of low order models in gaining a qualitative understanding of phonation. However quantitative discrepancies remain large due to an inaccurate estimation of the model parameters and the crudeness in either flow or mechanical model description. As an illustration it is shown that significant improvements can be made by accounting for viscous flow effects.

Analyse dynamique de la réduction vocalique en contexte CV à partir des pentes formantiques en arabe dialectal et en français

Jalaleddin Al-Tamimi.

Linear regression parameters (formant slopes and intercepts) were proposed to measure the degree of vowel reduction in 3 vowel systems: Moroccan Arabic, Jordanian Arabic and French. 10 speakers per language produced a list of vowels in C1VC, C1VCV or C1VCVC words, where C1 was /b/, /d/ or /k/. Our results show that the values of formant slope and intercept are dependent on: 1) the place of articulation of adjacent consonant, 2) the vowel quality, and 3) the language's vowel system density. Discriminant analysis results show the possibility of language separation on the basis of duration, F1 and F2 slope and intercept values.

SESSION ORALE 6 : Production et perception

Estimation des dyspériodicités vocales dans la parole connectée dysphonique

Abdellah Kacha, Francis Grenez, Jean Schoentgen.

Acoustic analysis of connected speech is carried out by means of a generalized variogram to extract vocal dysperiocities. A segmental signal-to-dysperiodicity ratio is used to summarize the perceived degree of hoarseness. The corpora comprise four French sentences as well as vowels [a] produced by 22 male and female normophonic and dysphonic speakers. It is shown that the segmental signal-to-dysperiodicity ratio correlates better with perceptual scores of hoarseness than the global signal-to-dysperiodicity ratio. The perceptual scores are based on comparative judgments by six listeners of pairs of speech stimulus.

L'intégration bimodale de l'anticipation du flux vocalique dans le flux consonantique

Emilie Troille, Marie-Agnès Cathiard.

It is well known that speech can be seen before it is heard: this has been repeatedly shown for the vowel rounding anticipatory gesture leading the sound (Cathiard [6]). In this study, the perception of French vowel [y] anticipatory coarticulation was tested throughout a voiced fricative consonant [z] with a gating paradigm. It was found that vowel auditory information, as carried by the noise of the fricative, was ahead of visual and even audiovisual information. Hence the time course of bimodal information in speech cannot be considered to display the same pattern whatever the timing of the coordination of speech gestures. As concerns vowel information only, consonantal coarticulation can carry earlier auditory information than the vowel itself, this depending of the structure of the stimulus. In our fricative-vowel case, it was obvious that the vowel building movement was audible throughout the fricative noise, whereas the changes in formant grouping occurred later.

Organisation syllabique dans des suites de consonnes en berbère : quelles évidences phonétiques?

Rachid Ridouane, Cécile Fougeron.

In this study, we examine consonants sequences in Tashlhiyt Berber in order to demonstrate that their syllabic organization can surface in their phonetic properties. Two types of three consonants sequences varying according to the degree of sonority of C1 are considered. Following syllabification principles of the language, these two types are considered to differ in their syllabic structure. Observation of the linguopalatal articulatory properties of the consonants and of the temporal coordination pattern between these consonants do show differences between the two types of sequences. These phonetic differences are interpreted as reflecting different syllabic structures, and results are confronted to the syllabification of the string proposed on phonological grounds.

Influence de la forme du palais sur la variabilité articulatoire

Jana Brunner, Pascal Perrier, Susanne Fuchs.

As has been noted previously, speakers with coronally low "flat" palates exhibit less articulatory variability than speakers with coronally high "domeshaped" palates. This phenomenon is investigated by means of a tongue model and an EPG experiment. The results show that acoustic variability depends on the shape of the vocal tract. The same articulatory variability leads to more acoustic variability if the palate is flat than if it is domeshaped. Furthermore, speakers with domeshaped palates show more articulatory variability than speakers with flat palates. The results are explained by different control strategies by the speakers. Speakers with flat palates reduce their articulatory variability in order to keep their acoustic variability low.

Peut-on parler sous l'eau avec un embout de détendeur ? Étude articulatoire et perceptive

Alain Ghio, Yohann Meynadier, Bernard Teston, Julie Locco, Sandrine Clairet.

We study the ability of sub aquatic divers to communicate by speech by means of an air regulator mouthpiece equipped with an acoustical sensor. These specific constraints on elocution led us to carry out an aerodynamic study to check phonation, an EPG study to observe the modification of articulation, and an analysis of labial forces involved with a special mouthpiece. Tests on intelligibility enabled us to evaluate the device in situation of real diving. In the current state, the various results let foresee a reduced but real possibility of spoken communication with a mouthpiece to certain conditions.

Production des voyelles nasales en français québécois

Véronique Delvaux.

This paper aims at describing the production of nasal vowels by 5 speakers of Canadian French (Montreal). The data consist in images of the tongue that have been tracked by ultra-sound while simultaneously recording nasal airflow with PcQuirer and the movements of the lips using a video camera. Results show that: (i) nasalization is delayed in Canadian French nasal vowels (especially in /e~/); (ii) the majority of the vowels are diphtongized, diphtongization being larger in front vowels than in back vowels and in closed syllables than in open syllables. The ways Canadian French deals with the constraints acting upon nasalization are discussed, especially in comparison with European French.

EXPOSÉ INVITÉ

De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux

Martine Adda-Decker.

Contrairement à bien d'autres domaines de recherche autour de la parole, la reconnaissance automatique, qui s'effectue sur un flux acoustique continu, nécessite une modélisation de l'ensemble des phénomènes observés dans le signal : au-delà des mots auxquels est associée une représentation de type phonologique dans le dictionnaire de prononciation, il faut modéliser des respirations, des hésitations, des fragments de mots, des brouillons de parole peu ou pas articulés... Dans cette intervention nous allons faire d'abord un état de l'art des systèmes de transcription automatique, présenter leurs performances et analyser les types d'erreurs de transcriptions les plus représentatifs. Nous allons ensuite poser la question de ce que peuvent nous apprendre ces erreurs de transcription. Ceci nous amène à utiliser progressivement les systèmes de transcription comme des instruments d'analyse de grands corpus oraux, permettant par exemple de décrire et de quantifier des variantes de prononciations, des disfluences, des réalisation acoustiques des sons. Quelques adaptations méthodologiques des systèmes s'imposent afin de transformer un système de transcription en un instrument d'analyse de corpus.

SESSION ORALE 7 : Corpus et variabilité

Détection automatique de frontières prosodiques dans la parole spontanée

Katarina Bartkova, Natalia Segal.

The present study addresses the issue of the automatic detection of prosodic units in French. An analysis of two prosodic parameters, phone duration and F0 slope values, carried out on two spontaneous speech data bases recorded by several thousands of speakers, revealed relevant deviations of these parameters on prosodic junctions. Vowel durations and F0 slopes are used to automatically detect prosodic units on the two data bases. The phone duration is modelled as the ratio of two subsequent vowel durations. Apart from the duration ratio, duration values are also modelled. The detection of prosodic units based on the F0 uses the value of the F0 slope and its standard deviation, recalculated after each pause. An evaluation of the automatic detection is carried out by comparing the prosodic border locations with the lexical boundaries and also with prosodic boundaries obtained by manual segmentation.

Analyses formantiques automatiques en français : périphéralité des voyelles orales en fonction de la position prosodique.

Cedric Gendrot, Martine Adda-Decker.

The aim of the present study is to investigate peripherality of French vowels in two prosodic positions: (i) word-final syllables (as compared to word-initial syllables), and (ii) in the vicinity (before and after) of pauses. The LIMSI speech alignment system is used and formant values of oral vowels are automatically measured in a total of 25000 segments from two hours of journalistic broadcast speech in French. A tendency to reduction for all vowels (in terms of the shrinking of the vocalic triangle formed by F1 and F2 values) of short duration was clearly observed in a former study (Gendrot & Adda-Decker [6]). We also attempt to check here whether the same relationship holds for vowels in both word-final and word-initial syllables.

Les systèmes vocaliques des dialectes de l'anglais britannique

Emmanuel Ferragne, François Pellegrino.

This paper is an attempt to characterize the vowel systems of the dialects of British English. We carried out a (semi-) automatic dialect identification procedure using [hVd] words. Our second aim was to examine to what extent the procedure allowed the description of vowel systems. The method yields approximately 90% correct identification, and we show that it is not sensitive to gender differences, and may therefore be used for the description of vowel systems.

SESSION POSTERS 5

Reconnaissance audiovisuelle de la parole par VMike

Fabian Brugger, Leila Zouari, Hervé Bredin, Asmaa Amehraye, Dominique Pastor.

This article presents a new Electronic Retina based Smart Microphone (VMike) and investigates the use of its novel parameters - lip profiles - in audiovisual speech recognition. In order to evaluate the parameterization, both an audio only and a video only speech recognition system are developed and tested. Then, two main fusion techniques are employed to test the usability of profiles in audiovisual systems: feature fusion and decision fusion. These results are compared to the performance of recognizers based on a state-of-the-art parameterization, and also to results obtained by applying perceptual filtering to the speech signal prior to recognition. When feature fusion is applied, and under noisy conditions, recognition using lip profiles improved by up to 13 percent with respect to audio-only recognition.

Probabilité a posteriori : amélioration d'une mesure de confiance en reconnaissance de la parole

Julie Mauclair, Yannick Estève, Paul Deléglise.

This paper adresses the word posterior probability used as a confidence measure on speech recognition system. We present a new confidence measure based on the behavior of language model back-off used during the recognition processing. Merging this new confidence measure with word posterior probability allows to obtain a fusion confidence measure, called WP/LMBB, which outperfoms the word posterior probability. Our experiments have been carried out on the corpus used during ESTER, the french evaluation campaign on automatic transcription of french broadcast news. Using the normalized cross entropy (NCE) as an evaluation metric, which is used in NIST evaluations, experimental results on test data of ESTER evaluation show a very significant improvement: whereas the word posterior probability reaches a value of NCE equal to 0.187, the WP/LMBB measure obtains 0.270.

Facteurs caractérisant les hésitations dans les grands corpus : langue, genre, style de parole et compétence linguistique.

Ioana Vasilescu, Martine Adda-Decker.

This paper deals with the factors characterizing the autonomous vocalic filled pauses in large spontaneous speech corpora, namely language, gender, speaking style and language proficiency. Two corpora are analyzed: a corpus of broadcast news in French and American English and a corpus of short talks in a conference in English spoken by native and non-native speakers. Several acoustic and prosodic parameters are evaluated and correlated with each factor, namely timbre, pitch, duration and density. Results presented here show that the timbre is correlated with language and language proficiency, whereas the duration is linked both to gender and speaking style, the latter conditioning also the hesitation density in speech.

Étude de disfluences dans un corpus linguistiquement contraint

Jean-Léon Bouraoui, Nadine Vigouroux.

This paper presents a study carried out on an air traffic control corpus which presents some specificity: apprenticeship situation, and the fact that the production is subordinate to a particular phraseology. Our study is related to the many kinds of disfluency phenomena that occur in this corpus, and the way they are or not affected by the nature of the corpus. We define 6 main categories of these phenomena. We then present the distribution of these categories. It appears that some of the occurrences frequencies largely differ from those observed in other studies. Our explanation is based on the corpus specificity: in reason of their responsibilities, both controllers and pseudo-pilots have to be especially careful to the mistakes they could do, since they could lead to some dramas.

Évolution de la perception des phonèmes, mots et phrases chez l'enfant avec implant cochléaire : un suivi de trois ans post-implant

Victoria Medina, Willy Serniclaes.

The aim of the present study was to examine the development of the perception of phonemes, words and sentences in a group of 18 children with cochlear implant (IC) from 12 to 36 months after implantation. The results show that the perceptual development of the different segments is fairly linear, that the rate of development is faster for words vs. phonemes and for sentences vs. words and that consonant perception, but not vowel perception, predicts later development of word and sentence perception.

Étude de la dysprosodie Parkinsonienne: Analyses acoustiques d'un schéma de type interrogatif

karine rigaldie, Jean Luc Nespoulous, nadine Vigouroux.

This article aims to acquire a better knowledge of prosodic disturbances in Parkinson disease via an acoustic analysis. The investigation of the patients' vocal productions by the way of acoustic analyses should indeed allow to identify phonetic and prosodic parameters that are specific of such a pathology. The Parkinsonian subjects had to repeat the interrogative pattern "Vous avez appris la nouvelle?" (in english: «You heard the news?»), three times: at the beginning, in the middle, and in the end of the protocol. This timing was determined in order to evaluate the effects of tiredness and the influence of other stimuli during the protocol. In order to determine the effect of dopamine, oral productions of twelve parkinsonian patients have been collected, in the OFF and ON states, and have then been compared to those of control subjects.

Corrélats auditifs et cognitifs à la capacité de restauration de la parole accélérée

Caroline Jacquier, Fanny Meunier.

We explore the relationship between auditory measures, reading capacities and the ability to reconstruct time-compressed speech for individuals without language trouble. We focused on two short attributes of speech: Voice Onset Time (VOT) and second formant transition. Normal hearing subjects had to identify disyllabic CVCV non-words that have been time-compressed on both acoustic cues simultaneously. The time compression experience showed a large inter-individual variance and allowed us to contrast a good performer and a bad performer groups for speech perception. Complementary studies (audiometric test and reading skills evaluation) showed that there is no correlation between auditory measures and cognitive mechanisms of degraded speech reconstruction whereas there are specifics correlations between reading capacities and performances in cognitive reconstruction.

Familiarité aux accents régionaux et identification de mots

Frédérique Girard, Caroline Floccia, Jeremy Goslin.

In this study on regional accent perception we conducted two experiments to examine the role of familiarity with a given regional accent upon the observation of a word identification cost. Participants were asked for a lexical decision on the last item of sentences uttered in a familiar or an unfamiliar regional accent. In the first experiment, a group of Besançon participants were presented with their home accent and a Toulouse accent. In the second experiment, a group of Toulouse participants were presented with the same stimuli, as well as with a list of Swiss French accented sentences. Results showed an interaction between participant groups and accent familiarity, suggesting that the word identification cost associated with a non-native regional accent can be predicted by participants' familiarity with this accent.

Nasalité consonantique et coarticulation : étude perceptive

Tiphaine Ouvaroff, Solange Rossato.

This paper investigates the coarticulation of consonantal nasality from a perceptive point of view. The aim of this study is to determine in CV utterances to which extent the consonant is perceived in the following vowel and compare these perceptual boundaries between oral and nasal consonants. Results show that, although the vowel is actually nasalized (low velum and consistent nasal air flow), the listeners don't attribute the nasalization of the vowel to the presence of a nasal consonant. After the release, the nasal feature of the consonant is lost, only the place of articulation is perceived until 60 ms after the release.

Reconnaissance automatique de phonèmes guidée par les syllabes
Olivier Le Blouch, Patrice Collen.

This paper presents a phonetic transcription system of French speech. This recognizer is based on a phonetic transcription driven by syllables, a syllabic bigram language modelling and a HMM topology adapted to syllables. The phone error rate obtained is compared to basic, usual systems at phonetic level : once the resulting syllables have been converted to phones, the phone error rate on a 12 minute-part of BREF80 corpus is as low as 15.8% with 35 phones.

Reconnaissance de parole non native fondée sur l'utilisation de confusion phonétique et de contraintes graphèmiques
Ghazi Bouselmi, Dominique Fohr, Irina Illina, Jean-Paul Haton.

This paper presents a fully automated approach for the recognition of non native speech based on acoustic model modification. For a native language (LM) and a spoken language (LP), pronunciation variants of the phones of LP are automatically extracted from an existing non native database. These variants are stored in a confusion matrix between phones of LP and sequences of phones of LM. This confusion concept deals with the problem of non existence of match between some LM and LP phones. The confusion matrix is then used to modify the acoustic models (HMMs) of LP phones by integrating corresponding LM phone models as alternative HMM paths. We introduce graphemic contraints in the confusion extraction process. We claim that prononciation errors may depend on the graphemes related to each phone. The modified ASR system achieved a significant improvement varying between 20.3% and 43.2% (relative) in «sentence error rate» and between 26.6% and 50.0% (relative) in «word error rate». The introduction of graphemic contraints in the phonetic confusion allowed improvements while using the word-loop grammar.

Étude par transillumination des consonnes occlusives simples et géminées de l'arabe marocain
Chakir Zeroual, Phil Hoole, Susanne Fuchs.

This study provides an acoustical and transillumination analysis of the laryngeal gestures responsible for VOT differences between voiceless plosives in Moroccan Arabic. The abduction and adduction phases and the interval between the maximal glottal opening and the release (MGO-REL) are longer during the geminate than during simple plosives. MGO-REL is shorter during the aspirated plosives than during unaspirated ones. Geminate plosives have a closure duration, a total duration and MGO that are larger than their simple correspondents. The closure duration is longer during unaspirated plosives than during aspirated ones. The voiceless geminate plosives have the same values of the VOT as their simple counterparts.

Analyse fibroscopique des consonnes sourdes en berbère
Rachid Ridouane.

This article deals with laryngeal adjustments during the production of singleton voiceless consonants in Tashlhiyt Berber. It focuses on the influence of place and manner of articulation and effects of position in the word. Results provide evidence that the degree of glottal opening as well as the velocity of abduction-adduction gestures vary according the place of articulation of stops and fricatives and their position in the word. Systematic differences, reflecting a universal tendency, were also observed between stops and fricatives. The specific laryngeal adjustments during the production of uvulars and so-called pharyngeals will be briefly outlined in the discussion.

Extraction des mouvements du conduit vocal à partir de données cinéradiographiques
Julie Fontecave, Frédéric Berthommier.

Since high speed X-ray films still provide the best dynamic view of the entire vocal tract, large existing databases have been preserved and are available for the speech research community. We propose a new technique for automatic extraction of vocal tract movements from these data. At first, the method was developed for the extraction of tongue movements in Wioland recorded in Strasbourg in 1977. Then, the same technique was adapted to other articulators and other X-rays films, taking into account their specificities. Finally, a quantitative evaluation of the estimate error and a comparison with Thimm and Luettin (1999) are achieved.

- **SESSION ORALE 8 : Analyse, codage et synthèse**

Bases théoriques et expérimentales pour une nouvelle méthode de séparation des composantes pseudo-harmoniques et bruitées de la parole
Laurent Girin.

In this paper, the problem of separating the harmonic and noise components of speech signals is addressed. A new method is proposed, based on two specific processes dedicated to better take into account the non-stationarity of speech signals: first, a period-scaled synchronous analysis of spectral parameters (amplitudes and phases) is done, referring to the Fourier series expansion of the signal, as opposed here to the typically used Short-Term Fourier Transform (STFT). Second, the separation itself is based on a low-pass filtering of the parameters trajectory. Preliminary experiments on synthetic speech show that the proposed method has the potential to significantly outperform a reference method based on STFT: Signal-to-error ratio gains of 5 dB are typically obtained. Conditions to go beyond the theoretical framework towards more practical applications on real speech signals are discussed.

Adjonction de contraintes visuelles pour l'inversion acoustique-articulatoire
Blaise Potard, Yves Laprie.

The goal of this work is to investigate audiovisual-to-articulatory inversion. It is well established that acoustic-to-articulatory inversion is an under-determined problem. On the other hand, there is strong evidence that human speakers/listeners exploit the multimodality of speech, and more particularly the articulatory cues: the view of visible articulators, i.e. jaw and lips, improves speech intelligibility. It is thus interesting to add constraints provided by the direct visual observation of the speaker's face. Visible data were obtained by stereo-vision and enable the 3D recovery of jaw and lip movements. These data were processed to fit the nature of parameters of Maeda's articulatory model. Inversion experiments show that constraints on visible articulatory parameters enable relevant articulatory trajectories to be recovered and substantially reduce time required to explore the articulatory codebook.

Estimation des instants de fermeture basée sur un coût d'adéquation du modèle LF à la source glottique
Damien Vincent, Olivier Rosec, Thierry Chonavel.

An algorithm for GCI (Glottal Closure Instants) estimation is presented in this paper. It relies on a source-filter model of speech production using a LF model for the source component. From this source-filter decomposition, a ratio which measures the goodness of fit of the LF source model is introduced in the GCI estimation procedure together with fundamental frequency constraints. Then, a Viterbi algorithm is applied to extract the most likely GCI sequence. Experiments performed on a real speech database show that the proposed method outperforms existing approaches.

Codage à bas débit des paramètres LSF par quantification vectorielle codée par treillis
Merouane Bouzid, Amar Djeradi, Bachir Boudraa.

Speech coders operating at low bit rates necessitate efficient encoding of the linear predictive coding (LPC) coefficients. Line spectral Frequencies (LSF) parameters are currently one of the most efficient choices of transmission parameters for the LPC coefficients. In this paper, an optimized trellis coded vector quantization (TCVQ) scheme for encoding the LSF parameters is developed. When the selection of a proper distortion measure is the most important issue in the design and operation of the encoder, an appropriate weighted distance measure has been used during the TCVQ construction process. Using this distance, we will show that our LSF TCVQ encoder performs better than the encoder conceived with the unweighted distance.

Évaluation d'un système de synthèse 3D de langue française parlée complétée
Guillaume Gibert, Gérard Bailly, Frédéric Elisei.

This paper presents the virtual speech cuer built in the context of the ARTUS project aiming at watermarking hand and face gestures of a virtual animated agent in a broadcasted audiovisual sequence. For deaf televiewers that master cued speech, the animated agent can be then incrusted - on demand and at the reception - in the original broadcast as an alternative to subtitling (as illustrated by Figure 1). The paper presents the multimodal text-to-speech synthesis system and the first evaluation performed by deaf users.

Modélisation B-spline de contours mélodiques avec estimation du nombre de paramètres libres par un critère MDL

Damien Lolive, Nelly Barbot, Olivier Boeffard.

This article describes a new approach to estimate F0 curves using a B-Spline model characterized by a knot sequence and associated control points. The free parameters of the model are the number of knots and their location. The free-knot placement, which is a NP-hard problem, is done using a global MLE within a simulated-annealing strategy. The optimal knots number estimation is provided by MDL methodology. Two criteria are proposed considering control points as real coefficients with variable precision. They differ on the precision used. Experiments are conducted in a speech processing context on a 7000 syllables french corpus. We show that a variable precision criterion gives good results in terms of RMS error (0.42Hz) as well as in terms of B-spline freedom number reduction (63% of the full model).

SESSION POSTERS 6

Constitution d'un corpus textuel basée sur la divergence de Kullback-Leibler pour la synthèse par corpus

Aleksandra Krul, Géraldine Damnati, Thierry Moudenc, François Yvon.

This paper presents a text design method for Text-To-Speech synthesis application. The aim of this method is to build a corpus whose unit distribution is close to a target distribution. As text selection is a NP-hard set covering problem, a greedy algorithm is used. We propose the Kullback-Leibler divergence to compute the score of each candidate sentence. The proposed criterion gives the posibility to control the unit distribution at each step of the algorithm. Finally, we present the first results and we compare the proposed criterion with two standard criteria.

eLite : système de synthèse de la parole à orientation linguistique

Richard Beaufort, Alain Ruelle.

eLite is the Text-to-Speech synthesis system developed by the TTS-NLP group of Multitel ASBL. The creation of eLite has been an opportunity for the group to carry out and integrate further research on all domains of text-to-speech synthesis, like morphological analysis, syntactic desambiguation and non-uniform units selection. This paper presents the general features and techniques of the system.

Coopération entre méthodes locales et globales pour la segmentation automatique de corpus dédiés à la synthèse vocale

Safaa Jarifi, Olivier Rosec, Dominique Pastor.

This paper introduces a new approach for the automatic segmentation of corpora dedicated to speech synthesis. The main idea behind this approach is to merge the outputs of three segmentation algorithms. The first one is the standard HMM-based (Hidden Markov Model) approach. The second algorithm uses a phone boundaries model, namely a GMM (Gaussian Mixture Model). The third method is based on Brandt's GLR (Generalized Likelihood Ratio) and aims to detect signal discontinuities in the vicinity of the HMM boundaries. Different fusion strategies are considered for each phonetic class. The experiments presented in this paper show that the proposed approach yields better accuracy than existing methods.

Influence des paramètres psycholinguistiques du cocktail party sur la compréhension d'un signal de parole cible

Claire Grataloup, Michel Hoen, Francois Pellegrino, Fanny Meunier.

This paper presents results from an experiment studying the cognitive ability to understand a speech signal in a babble background noise. We further tested subject's sensitivity to characteristics of target and competitor words. Our results show that words are better reconstructed than pseudowords. Intelligibility of words is not influenced by a change (number of voices, frequency of words) in the background babble noise whereas intelligibility of pseudowords is. Pseudowords perception is easier when words that constitute the background noise are low frequency words and when the number of voices is fewer.

Analyse des stratégies de chunking en interprétation simultanée

Myriam Piccaluga, Bernard Harmegnies.

In this paper, which is meant as a methodological account, we focus on a new variable (“Ecart Inter Syllabique”: EIS), intended to improve the study of speech chunks produced by subjects performing a task of simultaneous interpreting (“IS”). The variable is introduced on the basis of a discussion of the main methodological trends in the field, with the aim of improving the validity and reliability of the numerical treatments applied to the study of IS. An experimental essay is performed on a prototypical sample of 4 subjects, performing IS under several conditions. The behaviour of the variable within the design suggests its interest for future research.

Produit multiéchelle pour la détection des instants d’ouverture et de fermeture de la glotte sur le signal de parole

Aïcha Bouzid, Noureddine Ellouze.

This paper deals with robust singularity detection in speech signal using multiscale product method. These singularities correspond to opening and closure instants of the glottis (GOIs and GCIs). Multiscale product method consists of computing the products of wavelet transform coefficients of the speech signal at appropriate adjacent scales. As wavelet modulus maxima are a tool for signal edge detection, first derivative of a Gaussian function, is used for detecting speech signal discontinuities. Speech Multiscale products enhance edge detection. The proposed method is evaluated comparing to the EGG signal references using the Keele University database. This method gives excellent results concerning GOI and GCI detection from speech signal.

Modélisation 2D (« fréquence-temps ») des amplitudes spectrales

Mohammad Firouzmand, Laurent Girin.

This paper presents a method for modeling the spectral amplitude parameters of speech signals in “two dimensions” (2D). It consists in two cascaded modeling: the first one along the frequency axis is usual, since it consists in modeling the log-scaled spectral envelope with a sum of Discrete Cosine (DC) functions. The second one, along the time axis, consists in modeling the trajectory of the envelope DC parameters by another similar DC model. An iterative algorithm that optimally fits this 2D-model, taking into account perceptual criterions, is proposed. This approach is shown to provide an efficient representation of speech spectral amplitude parameters in terms of coefficient rates, while providing good signal quality, opening new perspectives in very-low bit-rate speech coding.

Réduction du débit des LSF par un système d'énumération en treillis

Bachir Boudraa, Malika Boudraa, Mouloud Djamah, Merouane Bouzid, Bernard Guerin.

In the present study, we were interested in the reduction of the bit-rate observed in the speech coder named CELP FS1016 (federal standard developed by the US department of the defense «DoD «). More precisely, the quantization of the Line Spectrum Frequencies (LSF) parameters was concerned. In the standard CELP FS1016, these coefficients are coded with 34 bits. We propose the use of an enumeration technique in conjunction with a treillis search coding schemes that exploits the natural ordering of the LSF. The technique allows reducing the bit rate of the LSF coefficients to 30 bits without decreasing the performance of the coder.

Évaluation de la qualité vocale dans les télécommunications

Marie Guéguin, Vincent Barriac, Valérie Gautier-Turbin, Régine Le Bouquin-Jeannès, Gérard Faucon.

This paper is a review of the methods for speech quality assessment. Subjective methods involve human subjects testing systems in various network conditions and voting on an opinion scale. The scores obtained for each condition are averaged to get a mean opinion score (MOS). These subjective tests are the only way to assess perceived speech quality, but they are complex, cost- and time-consuming. Consequently objective methods have been introduced to predict the speech quality as perceived by users. Here, objective methods are classified depending on the context they deal with. This review of objective methods shows a lack of model in the conversational context. Then we propose an objective model of the conversational speech quality, built on a combination of objective models of the listening and talking speech qualities and the delay.

La répétition stylistique en anglais oral

Gaëlle Ferré.

This paper is based on a video recording of a face to face interaction between two British girls. In another study on the characteristics of young people's speech involving the same corpus I noticed that one of the specificities of my two speakers lied in the constant repetitions of segments. Some segments are not only repeated in the case of hesitation but also as a stylistic device. I propose to describe in the present paper the stylistic repetitions in terms of what kind of segments are repeated and what is the role of such repetitions. Taking into account lexico-syntactic, prosodic and gestural parameters, I will also show that these repetitions cannot be assimilated to some hesitation on the parts of the speakers.

Cohésion temporelle dans les groupes C1/l/ initiaux en français

Barbara Kühnert, Phil Hoole.

This work examines aspects of inter-consonantal cohesion within French word-initial C1/l/-clusters in light of recent proposals of gestural coordination. Based on articulatory and acoustic events, the timing of tongue and lip movements in one subject was studied using an electromagnetic transduction device. More temporal overlap between C1 and /l/ gesture onset as well as /l/ closure period was found for /pl/ than /kl/. Although matching similar patterns of overlap in initial stop clusters, this 'place of articulation' effect is attributed to low-level motor factors rather than to considerations of perceptual recoverability. An additional analysis of the overall C-centre of /pl/ and /kl/ showed a surprising temporal stability, confirming a relative constant phasing between initial consonant sequence and following vowel.

Étude des adductions/abductions totales et partielles des cordes vocales

Chakir Zeroual, John H. Esling, Lise Crevier-Buchman.

In this study, we have shown that, in the intervocalic position, simple and geminate voiceless plosives [t tt], present a total abduction of the vocal folds (anterior+posterior parts). While their voiceless emphatic (or uvularized) counterparts [T TT] are produced with an anterior abduction only. Simple and geminate voiced plosives also show a slight anterior abduction which is longer during geminate, and shorter during [gg]. Arguments have been presented showing that the anterior abductions observed during [T TT] and voiced plosives are passive, due to the increase of the intraoral pressure. [i] falsetto shows a slight abduction of the vocal folds that we assign to the action of the intrinsic laryngeal muscles.

EXPOSE INVITE

Imagerie cérébrale et apprentissage des langues

Christophe PALLIER

The neurolinguistics of bilinguism and language acquisi-tion is still in infancy. This paper presents a quick overview of some brain imaging studies of language acquisi-tion conducted in our lab. We describe a study showing that, as grammatical skills in L2 increase, the cerebral ac-tivations elicited by sentence building in L1 and in L2 became more and more similar. In another series of studies,we discovered anatomical and functional cerebral correlates of the abilities to memorize, perceive or produce foreign speech sounds.

SESSION ORALE 9 : Psycholinguistique, cognition, acquisition

Les effets de compétition lors de la reconnaissance des mots parlés : quand l'inhibition bottom-up joue un rôle.

Sophie Dufour, Ronald Peereman.

In two experiments, we examined lexical competition effects using the phonological priming paradigm in a repetition task. Experiment 1 showed that inhibitory priming effect occurs when the primes mismatched the targets on the last phoneme (/bagaR/–/bagaj/). In contrast, a facilitatory priming effect was observed when the primes mismatched the targets on the medial phoneme (/viRaj/-vilaj/). Experiment 2 replicated theses findings with primes presented visually rather than auditorily. The data thus indicate that the position of the mismatching phoneme is a critical factor in determining the competition effect in phonological priming. Such an observation suggests that both bottom-up inhibition and lexical competition are involved in the word recognition process.

L'émergence du contrôle segmental au stade du babillage : une étude acoustique
Mélanie Canault, Pascal Perrier, Rudolph Sock.

The aim of this work is to look for evidence that a segmental control of speech production could emerge during babbling from mandible rhythm dominance. Our assumption is that it could be found in temporal modulations of mandibular cycle phases . Acoustic analyses of two subjects between 10 and 15 months of age reveal that at 10 months, the temporal patterns of these children' productions are variable, before becoming at 15 months more stable and similar to adults' temporal patterns. These findings are interpreted as consequences of the emergence of a speech specific segmental control guided by the imitation of adults production.

L'implication des contraintes motrices dans "l'effet Labial Coronal"
Amélie Rochet-Capellan, Jean-luc Schwartz.

Stability of LC (Labial-Coronal) and CL CVCV se-quences was compared using the paradigm of reiterant speech with rate increase. The rationale was that rate in-crease would lead the articulatory system towards its most stable coordination mode . A first study analyzed the acoustic productions of 28 French speakers. Then, a sec-ond study focused on the articulatory coordination for 5 speakers. Results show that the repetition of LC and CL CVCV sequences could both evolve towards an LC (/pata/->/patá->/ptá) or a CL (/pata/->/páta->/tpá) attrac-tor. Yet, the LC attractor is largely favored compared with the CL one. Moreover, rate increase drives lips and tongue occlusions close together on a single jaw cycle. These results provide new elements to explain the LC effect in world languages by motor control constraints.

Stratégie de segmentation prosodique : rôle des proéminences initiales et finales dans l'acquisition d'une langue artificielle
Odile Bagou, Ulrich H. Frauenfelder.

Language acquisition in infants and adults depends upon both the segmentation of the words in the speech chain and the extraction of language-specific regularities, particularly prosodic regularities. Two experiments investigate whether prosodic information provided by accented syllables located at the beginning and at the end of prosodic words is used by adult French learners to segment an artificial language. The results allow us to define a prosodic strategy of segmentation: (1) prominence in word final position are used to hypothesize final boundaries; and (2) cues in word initial position can be used if and only if a primary final prominence is present.

Tomber le masque de l'information : effet cocktail party, masque nformationnel et interférences psycholinguistiques en situation de compréhension de la parole dans la parole
Michel Hoen, Claire-Léonie Grataloup, Nicolas Grimault, Fabien Perrin, François Pellegrino.

Up to now speech in noise comprehension and more particularly speech in concurrent speech sounds was rarely studied in the domain of psycholinguistics. In this paper we report a study interested in the differential effects of different types of speech derived noises as multi-talker cocktail party sounds and their time-reversed pendent on the comprehension of isolated words. Results suggest that different levels of linguistic information from concurrent speech signals can compete with linguistic information in the target signal, mainly depending on the spectral saturation caused by the increasing number of voices in concurrent signals. These results suggest linguistically specific participations in informational masking effects occurring in the context of speech in speech comprehension.

NOTES

NOTES

Avec le soutien de :