UMR IRISA

Vers plus de contrôle pour la synthèse de parole expressive

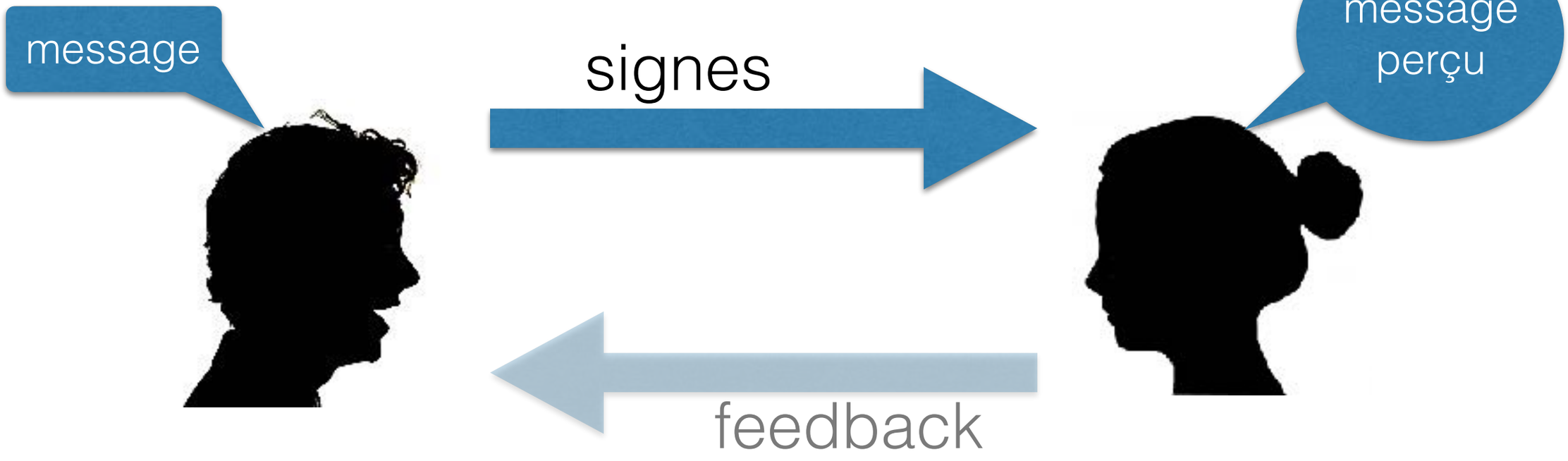
Damien Lolive
Enssat, Univ Rennes, IRISA, Expression

29 novembre 2017

Véronique Delvaux — Chercheuse qualifiée FNRS, Université de Mons
Frédéric Béchet — Professeur, Aix Marseille Université
Yves Laprie — Directeur de recherche CNRS, LORIA
Philippe Martin — Professeur, Université Paris Diderot
François Goasdoué — Professeur, Université de Rennes 1

Institut de Recherche en Informatique et Systèmes Aléatoires

Communication / Interaction



Communication / Interaction

- **Communication verbale / non verbale**
  - Transmettre un message sans le verbaliser
  - Postures, gestes, expressions faciales, regards
- **Systèmes et applications actuels**
  - Focus sur l’interaction textuelle
  - Reconnaissance vocale
  - Synthèse de parole pour applications spécifiques
- **Améliorations nécessaires sur chaque modalité pour favoriser l’interaction**

La parole

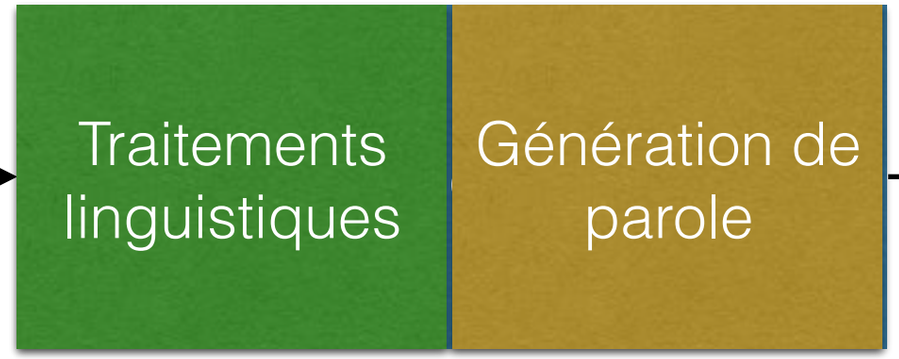
- **Parole = moyen de communication**
- Modalité par défaut utilisée par les humains pour communiquer en utilisant le langage
- Débit de transmission d'information élevé
- Evolution de l'homme : renforcement de la robustesse de la parole
  - amélioration des articulateurs (e.g. langue, larynx),
  - évolution du cerveau, capacités cognitives
  - contrôle moteur de la parole

Synthèse de parole et applications

Génération d'un signal de parole correspondant à une entrée donnée (souvent un texte)

La synthèse de parole

Je suis une voix de synthèse ***+ style / émotion / expressivité***



Normalisation du texte,
Phonétisation,
Syllabation,
Prédiction de la prosodie

Sélection de segments,
ou de modèles,
génération du signal

Parole et expressivité

Problématique

- **Adaptation de la parole au contexte :**
  - Selon les usages : GPS *vs.* jeux vidéo *vs.* style direct
  - Selon les utilisateurs : préférences
  - Selon la voix utilisée : prise en compte des caractéristiques du locuteur, i.e. sexe, accent, origine sociale, etc.

➡ **Meilleur contrôle de l'expressivité nécessaire**

➡ **Prise en compte à tous les niveaux**

Contributions

Librairie roots

Dépôt APP

PostDoc J. Chevelu

Corpus

Méthodologies d'évaluation

PostDoc J. Chevelu

ANR Phorevox

Je suis une voix de synthèse ***+ style / émotion / expressivité***

Traitements linguistiques

Génération de parole

Modélisation de la prononciation

PhD R. Qader, PostDoc M. Tahon

ANR SynPaFlex

Styles de parole

PhD A. Sini, PostDoc M. Tahon

ANR Phorevox, ANR SynPaFlex

Moteur de synthèse de parole et fonctions de coût

PhD D. Guennec

ANR SynPaFlex

Plan de la présentation

- Adaptation du système au…
  - Locuteur
  - Style
  - Public
- Moteur de synthèse de parole
  - Contrôle du processus de synthèse
  - Evaluation et cas d’usage
- Bilan et perspectives

Plan de la présentation

- **Adaptation du système au…**
  - **Locuteur**
  - **Style**
  - **Public**
- Moteur de synthèse de parole
  - Contrôle du processus de synthèse
  - Evaluation et cas d’usage
- Bilan et perspectives

Adaptation au locuteur : cas de la prononciation

- **Objectif :** [ANR SynPaFlex, 2015-2019, Post-doc M. Tahon]
  - Modifier la prononciation en fonction du locuteur
  - Rendre la parole synthétique plus expressive et plus *personnelle*
  - Augmenter la cohérence entre requête et contenu du corpus
- **Méthodologie :**
  - Champs conditionnels aléatoires - CRF
  - Modèle séquence (référence) vers séquence (réalisé)

Adaptation au locuteur : cas de la prononciation

- Exemple : « la guerre devient un peu moins improbable » [Interspeech 2016, SLSP 2016] *[Journal AFF. COMP.]*

Modèle	Séquence de phonèmes
Sans adaptation	l a g ɛ ʁ ə d ø v j ɛ̃ - ɛ̃ p ø m w ɛ̃ - ɛ̃ p ʁ ɔ b a b l ə
Adapté (C)	l a g ɛ ʁ - d ø v j ɛ̃ - ɛ̃ p ø m w ɛ̃ - ɛ̃ p ʁ ɔ b a b l -
Adapté (CLPh)	l a g ɛ ʁ - d ø v j ɛ̃ - œ̃ p ø m w ɛ̃ - ɛ̃ p ʁ ɔ b a b l -
Adapté (CLPhPr)	l a g ɛ ʁ - d ø v j ɛ̃ - œ̃ p ø m w ɛ̃ - ɛ̃ p ʁ ɔ b a b l -
Réalisé	l a g ɛ ʁ - d ø v j ɛ̃ t œ̃ p ø m w ɛ̃ z ɛ̃ p ʁ ɔ b a b l -

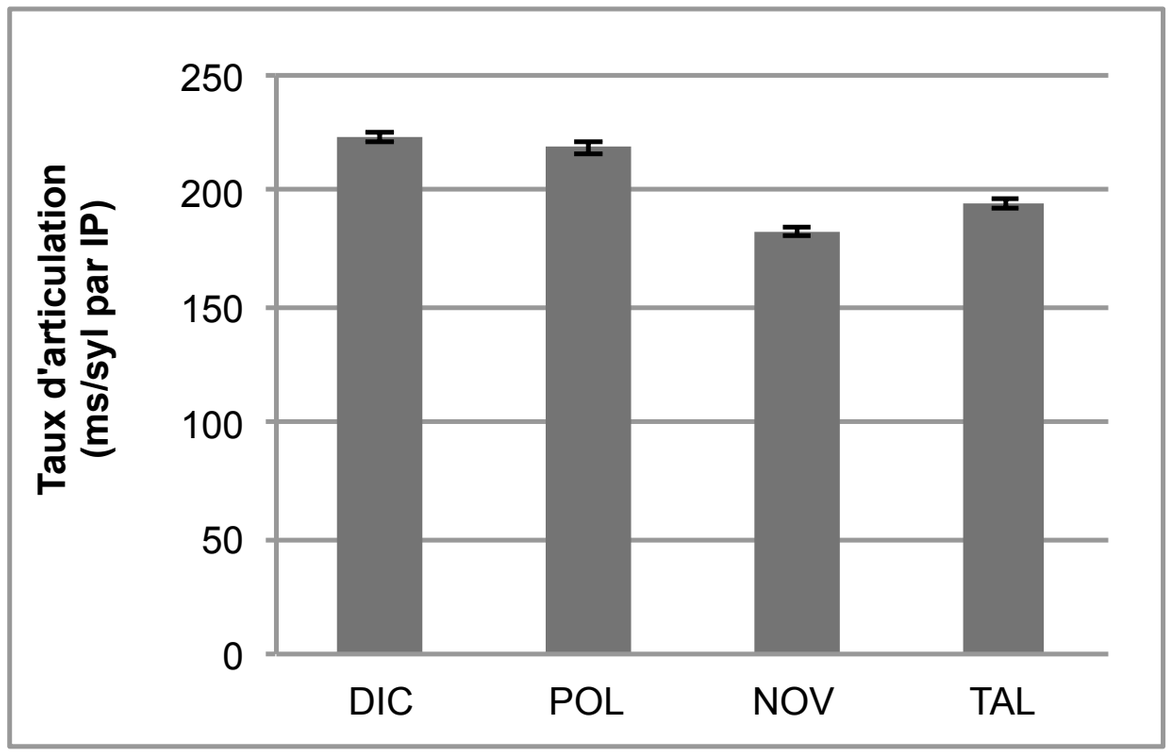
- Amélioration de la qualité de la parole générée
  - Plus de choix de segments, meilleures concaténations
  - Modèles contextuels mieux adaptés
  - Résultats probants avec peu de données

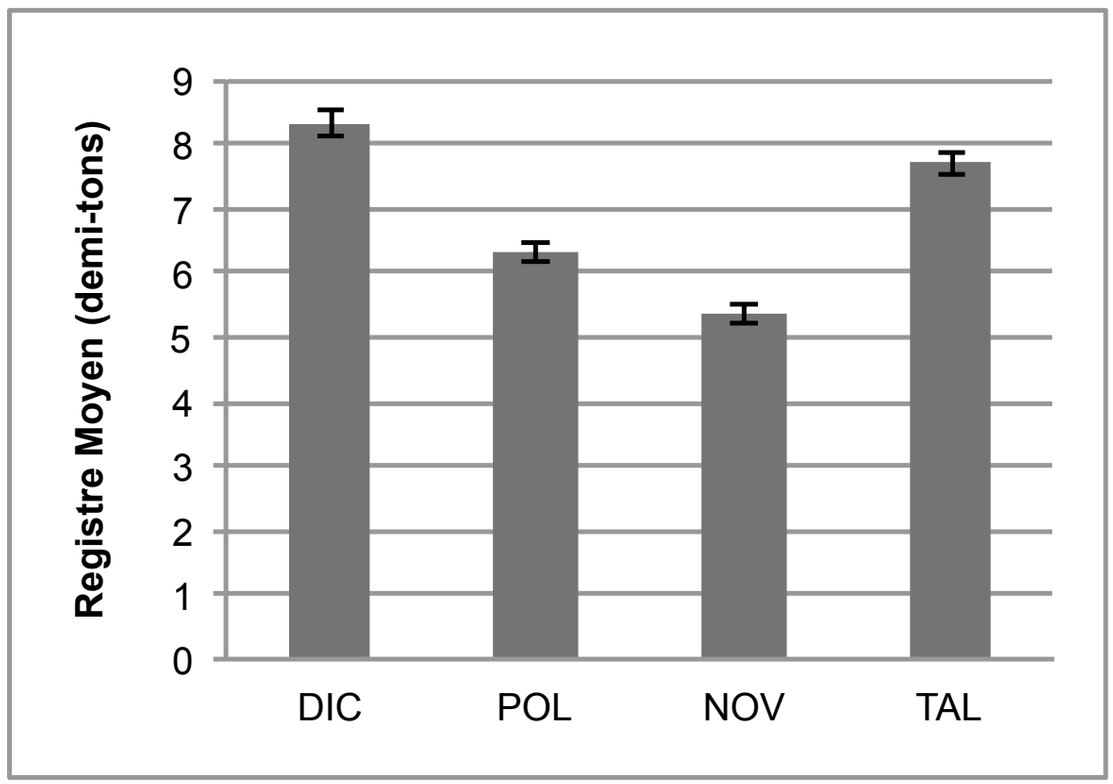
Adaptation au style : parole spontanée

- **Parole spontanée vs. Prononciation canonique** [Thèse, R. Qader, 2017]
  - ***~30% de phonèmes diffèrent, soit ~60% des mots***
  - Présence de disfluences : faux départs, répétitions, pauses
  - Contribue au naturel de la parole
  - Apporte des informations sur le locuteur
- **Contributions :** [SLSP 2015, JEP 2016, TSD 2017]
  - Adaptation de la prononciation vers spontané (anglais) *[Comp. Speech Lang.]*
  - Version adaptée plus naturelle mais moins intelligible
  - Prédiction des disfluences [TALN 2017] (Best paper award)
- **Difficultés :**
  - Evaluation de la parole spontanée ? Qualité de la synthèse ?

Adaptation au style : comparaison de différents genres

- **Analyse de différents styles :** dictée (DIC), discours politique (POL), roman (NOV), contes (TAL)





- **Différences significatives entre les styles**
- Exemple : dictée
  - pauses ↗, registre ↗, longueur AP/IP ↘

[Interspeech 2014]

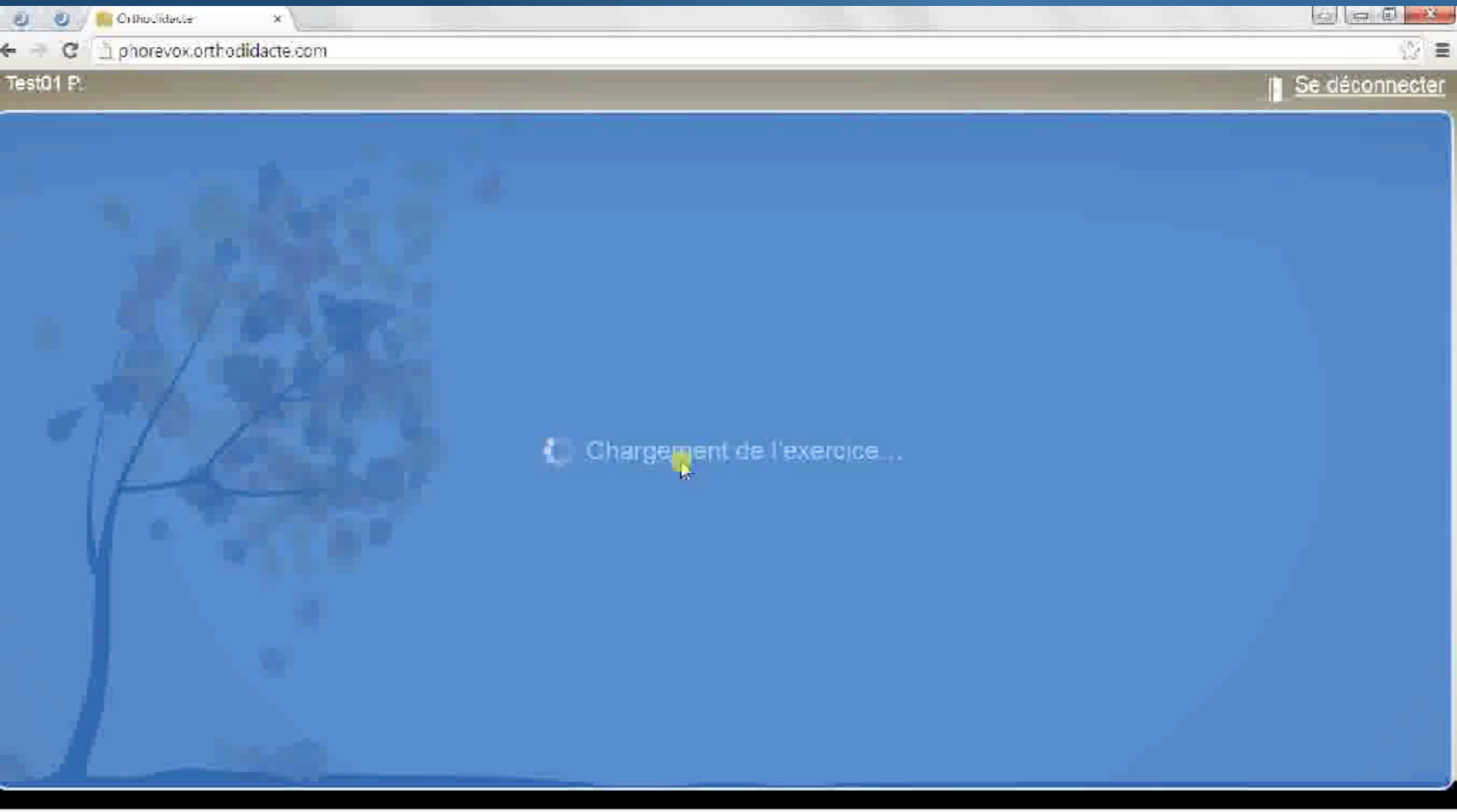
Adaptation au public

- **Contexte : apprentissage de l'écriture**
  - Public niveau cycle 2
  - Cas de la dictée
  - Prédiction des groupes prosodiques

[ANR Phorevox, 2012-2014]
[Interspeech 2014b]
[Speech Prosody 2014]
[JEP 2014]

- **Contribution :**
  - Algorithme de chunking adapté à la dictée
  - Evaluation en situation : élèves et enseignants
- **Résultats :**
  - Synthèse acceptée par les utilisateurs
  - Adaptation au style possible => nouvelles applications

Adaptation au public : Plateforme Phorevox



Plan de la présentation

- Adaptation du système au…
  - Locuteur
  - Style
  - Public
- **Moteur de synthèse de parole**
  - **Contrôle du processus de synthèse**
  - **Evaluation et cas d'usage**
- Bilan et perspectives

Synthèse par concaténation

- Développement système de synthèse IRISA depuis 2012
- Architecture classique : Filtres de pré-sélection + fonction de coût
- Mais écriture modulaire, langage moderne (C++)
- Recherche de la meilleure séquence de segments

PhD D. Guennec
[TSD 2014, JEP 2014]

- Points sensibles :
  - Jonctions entre segments
  - Segments sélectionnés v-à-v. cible attendue

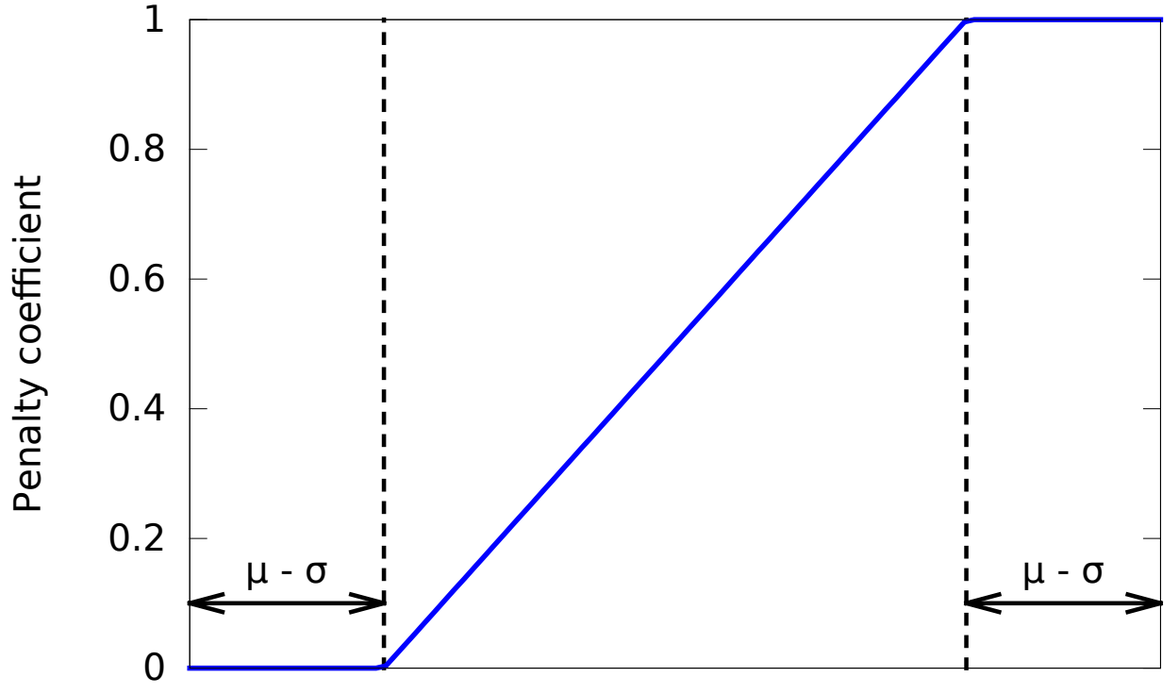
	Audiobook	*IVS*
Naturel	4.82 ±0.08	4.88 ±0.07
Système IRISA	3.38 ±0.25	3.17 ±0.21
Corpus bi-gram	2.14 ±0.14	1.72 ±0.08

Evaluation perceptive MOS

Améliorer les jonctions

- Favoriser/Eviter les jonctions sur certaines classes de phonèmes
- Origine : préparation de script d'enregistrement
- Expression rationnelle :

R(A*V A*)+R [Cadic et al., 2009]



- Pénalité fixe suivant la classe : $P_V \gg P_A > P_R$
- Pénalité pondérée par une fonction dépendant de la distribution des coûts

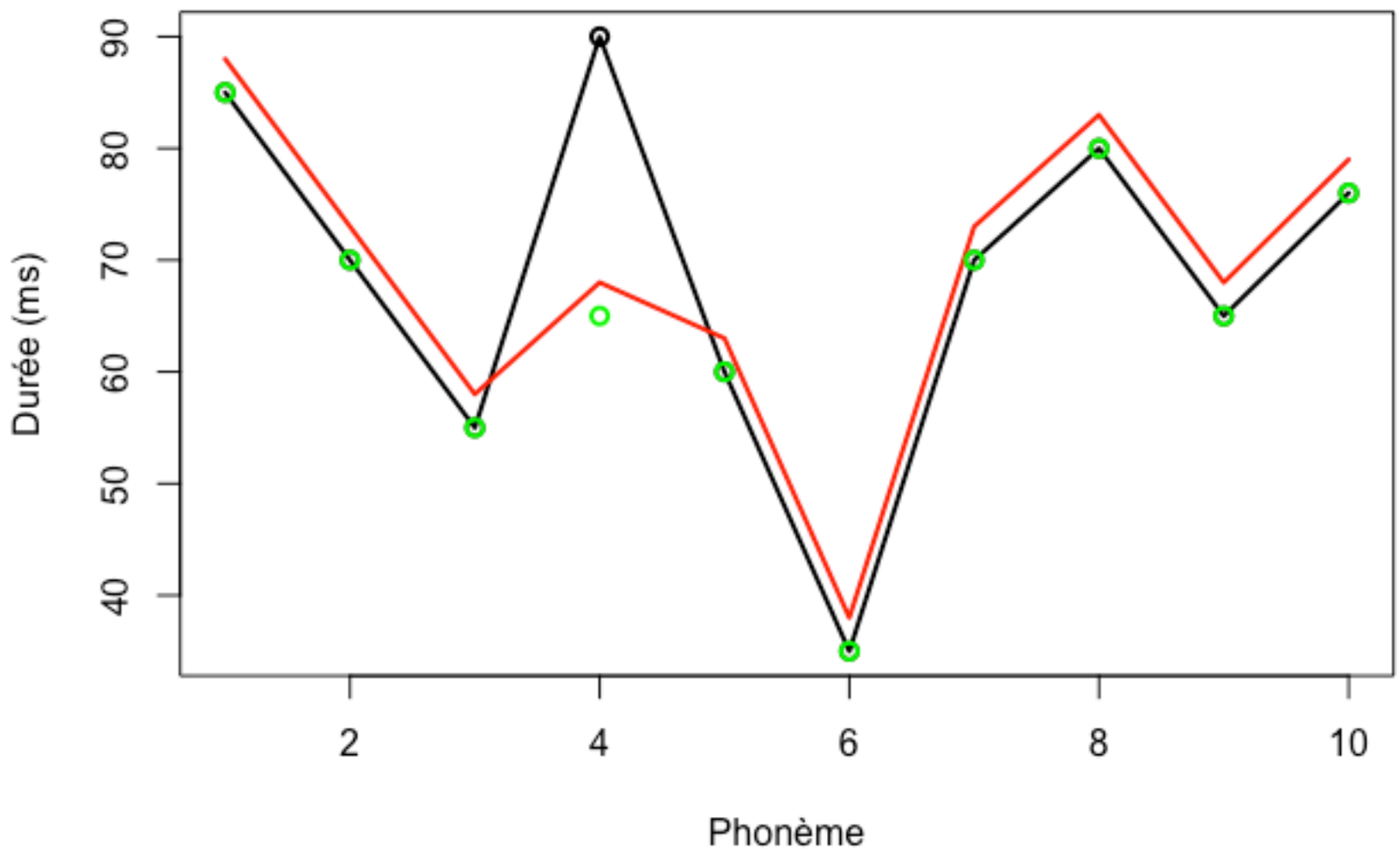
➡ Présence d'artéfacts ↘ mais nombre de concaténations ↗

[Interspeech 2016]
PhD D. Guennec

Mieux choisir les unités : cas de la durée

- Cas général :
  - Modèle pour prédire la durée + coût local
  - Apparition d'erreurs locales importantes

- **Idée :** préférer une modification globale des durées
- Prendre en compte l'erreur moyenne passée dans les choix locaux



[TSD 2015]

PhD D. Guennec

Evaluation

ANR Phorevox

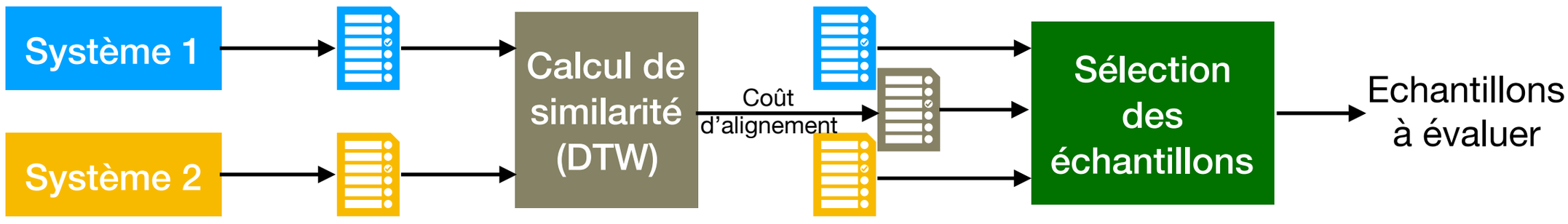
- Evaluation par l'humain **indispensable**!
  - Mais **difficile** : testeurs, question posée, volume d'échantillons, représentation statistique, contexte d'évaluation
    - 2 stratégies :
      - Evaluations « individuelles »
      - ou partagées : ***challenge Blizzard***
- Axe « représentativité des échantillons »
  - Choix aléatoire des échantillons à évaluer (souvent)
  - Autre méthode ?

Evaluation

- Focus sur les différences entre systèmes

ANR Phorevox
[Interspeech 2015]
[Eusipco 2015]

  - Ne pas évaluer plus d'échantillons
  - Mais choisir les échantillons « intéressants » à évaluer
  - Mesure objective de similarité comme critère de choix



- Résultats : significativité ↗
- Applicable aux tests avec 2 systèmes
- Extension à N systèmes ?

Challenge Blizzard

- **Challenge** international de synthèse de parole
  - ~10-12 participants chaque année
  - IRISA : depuis 2015

[Blizzard 2015]

- 2015 : Synthèse de 6 langues indiennes (Hindi, Bengali, etc.)
  - Script différent,

    बुज़ुर्ग कहा करते हैं , न दामाद के साथ खाओ , पुत्र के साथ पढ़ो

  - Chaîne de traitement à adapter
- Résultats : comparable aux autres systèmes
  - Connaissance sur les langues utilisées…
  - Et le volume de données (entre 2h et 4h / langue)
  - Evaluation difficile pour testeurs (par ex. WER)

Challenge Blizzard

[Blizzard 2016, 2017]

- 2016-2017 : Livres audio en anglais pour enfants :
  - Très expressifs, courts, nombreux noms propres, imitations, etc.
- Propositions :
  - Distinction Narration / style direct
  - Consigne prosodique : F0 moyen, taux d’articulation
- Nombreux critères d’évaluation :
  - impression globale, accentuation, pauses, intonation, émotion, effort d’écoute, etc.

Plan de la présentation

- Adaptation du système au…
  - Locuteur
  - Style
  - Public
- Moteur de synthèse de parole
  - Contrôle du processus de synthèse
  - Evaluation et cas d’usage
- **Bilan et perspectives**

Bilan : production scientifique

- **Publications**
  - 1 article international (+2 en cours de relecture) + 1 national
  - 32 conférences internationales + 15 conférences francophones
  - 2 dépôts logiciels à l'APP
- **Supervision**
  - 2 thèses soutenues (2016, 2017), 4 en cours (2015, 2016, 2*2017)
  - 2 post-doc
  - 4 masters, 1 en cours
- **Projets collaboratifs :** 2 projets ANR
  - Phorevox (PRCE) : apprentissage de la lecture
  - SynPaFlex (JCJC) : flexibilité de la synthèse de parole

Bilan et perspectives

- **Variantes de prononciation et disfluences**
  - ☑ Premiers travaux utilisés en synthèse de parole
  - ☑ Résultats encourageants
  - ☐ Spontané : diminution de la qualité de la synthèse
  - ☐ *Combiner modèle expressif et modèle du corpus ?*
  - ☐ *Reformulation sous contraintes (sémantique, expressivité) ?*
- **Adaptation de la prosodie**
  - ☑ Utilisation de la synthèse en contexte d'apprentissage des langues
  - ☑ Etude de différents styles -> règles applicables
  - ☐ *Autres facteurs ? qualité vocale ?*
  - ☐ *Modèles pour plus de flexibilité et de contrôle de la prosodie*

PhD A. Sini
Master 2

Bilan et perspectives

- **Synthèse de la parole**
  - ☑ Développement d'un moteur de synthèse
  - ☑ Intégration de contraintes phonologiques / prosodiques
  - ☑ Challenge Blizzard
    - ☐ *Meilleure représentation de l'espace acoustique nécessaire* — PhD A. Perquin
    - ☐ *Utilisation de données hétérogènes ?* — PhD C. Fayet
    - ☐ *Adaptation automatique de la voix au contexte ?* — PhD M. Shamsi
- **Méthodologies d'évaluation**
  - ☑ Focus sur les différences entre systèmes
  - ☑ Efficace mais applicable à seulement 2 systèmes — *=> extension nécessaire*
  - ☐ *Evaluation en contexte ?* — Collaboration Saarbruck

Perspectives

- **A plus long terme…**
  - Evolution importante du domaine ces dernières années
  - Wavenet, Deep Voice, etc.
  - Générer une parole de qualité : plus vraiment un problème !

- **Diversifier les applications**
  - Besoin de plus de contrôle
  - Nécessité d’établir un lien entre les niveaux sémantiques et le signal de parole
  - Synthèse adaptée au contexte : public, usage, etc.

Vers plus de contrôle pour la synthèse de parole expressive

Merci de votre attention

— — —