Voice Strength Representation and Estimation from the Long Term Amplitude Spectrum

Jean-Sylvain Liénard

Représentation et Estimation de la Force de Voix à partir du Spectre Moyen à Long Terme

Jean-Sylvain Liénard LIMSI, CNRS, Université Paris-Saclay

Problématique de l'Effort Vocal

- EV: terme consacré, mais flou: 3 à 7 nuances, pas de définition objective
- lié à la situation: le parleur ajuste sa voix de façon à se faire "entendre" par l'auditeur qu'il vise ou imagine, malgré distance, bruit, pb d'audition...
- entraîne variabilité spectro-temporelle du signal oral, qui altère les mesures en acoustique phonétique et les performances en traitement automatique
- recherche d'indices: F0, F1, spectral tilt, durée et intensité relatives C et V, forme et largeur de l'impulsion glottique: travaux dispersés, peu cohérents
- mais cette variabilité est précisément ce qui permet à l'auditeur de reconnaître le degré d'EV utilisé par le parleur: information paralinguistique

PROPOSITION

→ indexer l'EV par l'intensité émise (Force de Voix, en dB)

→ montrer que la FDV est codée dans la structure spectro-temporelle du signal

- problème: manque de données calibrées → corpus de Pearsons 1977

Les données de Pearsons (1977)

● objectif: normes de bruit préservant l'intelligibilité en divers lieux: transports, écoles, hôpitaux, logements; enregistrements calibrés en niveau

● plus: un enregistrement de référence en chambre anéchoïque: 97 locuteurs: 48 m, 37 f, 12 enfants (< 13 ans)

● phrase *"Joe took father's shoe bench out; she was waiting at my lawn"* répétée pendant 10 à 20 s, selon 5 consignes vocales: *"normal"*, *"raised"*, *"loud"*, *"shout"* + *"casual"* (conversation informelle, vx faible) → 482 enregistrements représentant environ 2h de parole

● enregistrements sonores perdus, mais les 482 SMLT (Spectre Moyen à Long Terme) ont été réhabilités par Anthony Nash (2014)

Représentation des données SMLT

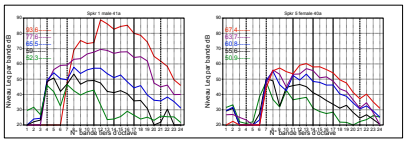
● **Modalités d'analyse**

● 24 canaux en tiers d'octave
- ● de 50 Hz (canal 1) à 10 kHz (canal 24)
- ● canal 14 (1 kHz) central → dilatation de la zone BF
- ● marques à 100, 200, 500 Hz, 1, 2, 5 kHz

● stabilisation en env. 40 s, ou 15 s à contenu constant

● niveau (L_{EQ}): intensité moyenne émise par seconde, mesurée à 1 m, exprimée en dB SP → force de voix (FDV)

● **deux exemples individuels**: locuteur 1 et locutrice 5:



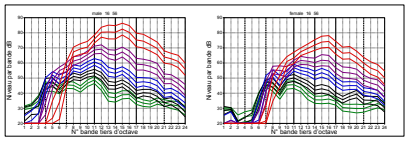
→ *la consigne vocale est diversement respectée, surtout en modes "loud" et "shout"*

→ *décalage 1 octave en BF entre voix m et voix f*

→ *canaux 1 et 2 altérés par bruit de fond*

→ **la non-linéarité des déformations spectrales explique la difficulté à définir des indices cohérents de l'Effort Vocal**

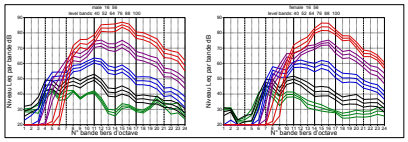
● **SMLT moyennés par consigne vocale**:
- ● Chaque SMLT moyen est entouré par deux bandes à ± 0,5 écart-type (→ 38% des observations)



→ *forte dispersion pour une même consigne vocale*

→ *en modes loud et shout: voix f moins fortes que voix m ?*

● **SMLT moyennés en 5 catégories de force de voix**:
- ● id, bandes à ± 0.5 écart-type



→ *dispersion moindre et plus grande dynamique*

→ *certaines voix f sont aussi fortes que les voix m criées*

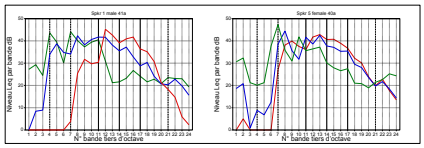
Normalisation et seuillage

● objectif: éliminer l'information de niveau (FDV) des SMLT , car c'est ce que l'on cherche à retrouver d'après leur profil spectral

→ **normalisation** à un niveau global arbitraire, p.ex. 50 dB

→ et **seuillage** à 0 dB pour réduire l'influence du recul de bruit de fond

● **exemples individuels**: locuteurs 1 et 5, modes casual, raised et shout, SMLT normalisés 0:50 dB



→ *les profils des SMLT de même FDV et même genre se ressemblent*

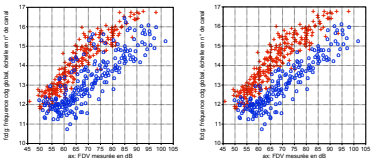
→ *rôle des canaux extrêmes: bruit de fond, prise de son*

Deux distributions m et f+e

● exemple: position fréquentielle du centre de gravité global du SMLT normalisé 0:50 dB, en fonction de la FDV

● voix des garçons plus proches des voix f, jusqu'à quel âge ?

● regroupement f + e<13 ans (à g), f + e<16 ans (à dr)

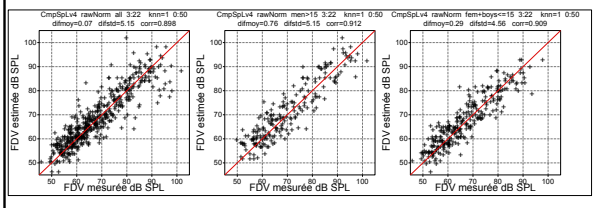


→ nuages m (en bleu) et f+e (en rouge) quasi-parallèles à env 2/3 d'octave à égale FDV, ou 10 à 15 dB à égale fréquence

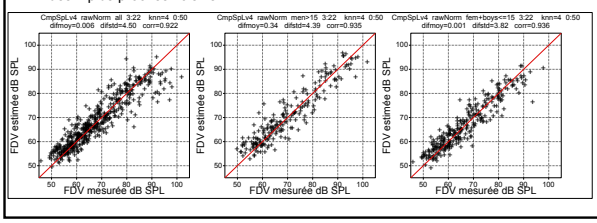
→ voix garçons < 16 ans se regroupent avec voix f

Estimation de la FDV par plus proches voisins spectraux

● **méthode du plus proche voisin** (ppv = 1):
1. comparer chaque SMLT normalisé à tous les autres, sauf ceux issus du même locuteur
2. FDV estimée = celle du plus proche voisin
3. représenter l'écart entre FDV estimée et FDV propre (éliminée par la normalisation)
4. performance évaluée par
   - Marge d'erreur statistique à 1 écart-type, en dB
   - Coefficient de corrélation
   - Forme du nuage



● **plusieurs voisins** (ppv = 4):
- Idem, mais la FDV estimée est la moyenne pondérée des FDV des 4 plus proches voisins



● **Corpus testés** (de g à dr)
- Corpus total m+f+e
- sous-corpus m
- sous-corpus f+e

● **Canaux** 3 à 22 (100 à 6300 Hz)

● **Normalisation** 0:50 dB

● **Performances**
- marges d'erreur env. 5 dB
- forte corrélation, env. 90%
- quelques erreurs grossières avec le corpus total, pour les fortes FDV

● **Remarques**
- peu d'erreurs grossières en corpus séparés m et f+e
- causes des erreurs grossières:
  - moindre densité des données en voix forte et criée
  - aspect réductif du SMLT
  - timbre propre à chaque parleur

● **Performances**
- marges d'erreur env. 4 dB
- très forte corrélation, env. 93%
- moins d'erreurs grossières

● **Remarque**
- réduit l'impact des erreurs dues à la granularité des données, mais n'en réduit pas la cause

Conclusions

● **Limites des données Pearsons**
- absence des enregistrements sonores
- manque de nuances en voix faible
- mode "casual" inhomogène avec les 4 autres modes
- prise de son à 1 m limite la dynamique des voix faibles

Mais: seules données calibrées d'envergure disponibles

● **Ce que montrent les résultats**
- la FDV est un descripteur objectif de l'EV, plus efficace que la consigne vocale
- la FDV peut être estimée à partir du SMLT en tiers d'octave
- éviter les canaux extrêmes (< 80 Hz et > 6 kHz), liés à la prise de son plus qu'à la FDV
- erreur < 5 dB et dynamique des voix > 50 dB: au moins 10 degrés de FDV objectivement codés dans le SMLT

● **Conséquences**
- la FDV est l'une des premières sources de variabilité acoustique, affectant:
  - la prosodie
  - les structures phonétiques de la parole
  - la caractérisation individuelle du parleur
- la FDV se manifeste par une variation de timbre: la FDV est une composante essentielle du timbre de la voix
- réhabiliter la dimension d'intensité

● **Perspectives**
- du SMLT à un nombre réduit d'indices spectro-temporels
- nécessité de bases de données calibrées
- acoustique phonétique: mesures à FDV comparable
- traitement automatique: intégrer estimation de la FDV