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

*G. Cloarec, D. Jouvet & J. Monné*

France Telecom – Division R&D – TECH/SSTP
2 avenue Pierre Marzin, 22307 LANNION, France
gwenael.cloarec@francetelecom.com

## ABSTRACT

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.

## 1. INTRODUCTION

Bien qu'étant des caractéristiques fondamentales du signal de parole, les paramètres relatifs à l'onde glottique, comme le pitch ou le voisement, ne sont que très rarement utilisés dans les systèmes de reconnaissance de la parole. Ceci est principalement dû au fait qu'ils sont considérés comme étant trop dépendants du locuteur. Cependant l'utilisation du paramètre de pitch est essentielle dans le cas des langues tonales, pour lesquelles le ton est utilisé pour la distinction lexicale. Le pitch peut par ailleurs être introduit pour améliorer la détection de fin de parole dans des conditions difficiles comme dans [1].

On peut néanmoins penser que l'utilisation de paramètres auxiliaires tels que le pitch pourrait améliorer le processus de reconnaissance en permettant, par exemple, de faciliter la distinction entre sons voisés et non voisés. En fait plusieurs études ont déjà été menées afin d'étudier les différents moyens d'intégrer ce coefficient au sein des systèmes de reconnaissance. La méthode la plus simple pour prendre en compte le pitch est de l'intégrer directement au sein du vecteur acoustique. Les résultats obtenus avec cette méthode montrent qu'il existe une différence de comportement entre les tâches de reconnaissance de mots isolés et celles de reconnaissance de parole continue. Ainsi dans [2], le pitch fut introduit au sein du vecteur acoustique et seule une légère amélioration fut obtenue sur de la reconnaissance de parole continue en Mandarin. De même, dans [3], l'utilisation du pitch n'amena qu'une faible amélioration sur de la reconnaissance de parole continue bien qu'une analyse linéaire discriminante ait été appliquée au vecteur acoustique afin de tirer profit de l'information apportée par le pitch, et dans [4] l'utilisation directe d'un indicateur de voisement ne conduisit qu'à de faibles améliorations, toujours sur de la reconnaissance de parole continue. Au contraire, dans [5] nous avons montré que l'utilisation du pitch et/ou du voisement pouvait améliorer significativement la reconnaissance de mots isolés.

Dans cet article nous étudions l'impact du pitch sur deux tâches de reconnaissance de natures différentes; une de reconnaissance de mots isolés, et une autre de reconnaissance de parole continue, et comparons les résultats obtenus.

Le papier est organisé de la façon suivante. La partie 2 décrit les conditions expérimentales. Dans la partie 3 nous reviendrons sur le paramètre de pitch ainsi que sur ses modélisations possibles. La partie 4 présentera les résultats expérimentaux obtenus ainsi que leur analyse. Finalement, les conclusions seront tirées dans la partie 5.

## 2. CONDITIONS EXPÉRIMENTALES

### 3.1. Généralités

Le système de reconnaissance utilisé est basé sur des modèles de Markov cachés. La modélisation acoustique prend en compte l'influence contextuelle des phonèmes et repose sur une modélisation gaussienne (8 gaussiennes par densité). L'analyse acoustique est réalisée par l'intermédiaire de l'algorithme Font-End ETSI ES 202 212. Les vecteurs acoustiques de référence sont composés de 33 coefficients : 10 coefficients mel-cepstraux (MFCCs) et le logarithme de l'énergie, ainsi que leurs dérivées temporelles première et seconde.

### 3.2 Bases de données

Les expériences ont été réalisées sur deux tâches de reconnaissance en mode indépendant du locuteur. La première, que nous appellerons *Communes* par la suite, est une tâche de reconnaissance de mots isolés. Elle est utilisée dans le cadre d'un annuaire téléphonique en environnement RTC, et est basée sur un vocabulaire d'environ 40 000 mots correspondant aux localités françaises [6]. Dans les résultats présentés par la suite, on n'utilise pas de modèle langage, tous les mots du vocabulaire sont donc équiprobables. La seconde, appelée *Plan Resto*, est une tâche de reconnaissance de parole

continue utilisée dans le cadre d'un service de renseignements touristiques avec le système de dialogue décrit dans [7]. Elle est basée sur un vocabulaire de 2200 mots et est utilisée en environnement RTC [8]. Le modèle de langage est basé sur un modèle bi-gramme

Le corpus de test de la tâche *Communes* est composé de 10571 données valides, c'est-à-dire de données correspondant au vocabulaire, et de 1635 données hors vocabulaire, donc à rejeter. Le corpus de test de la tâche *Plan Resto* comprend quant à lui 7507 phrases. Dans les 2 cas, il s'agit de parole spontanée recueillie dans le cadre d'expérimentation de services vocaux.

Les performances sont données en terme de *Word Accuracy*. Pour la tâche *Communes* les performances seront toutes données pour un taux de rejet à tort de 10%, afin de comparer les différentes modélisations du pitch pour un point de fonctionnement donné.

## 3. MODÉLISATION DU PITCH

### *3.1 Calcul du pitch*

Le pitch est calculé par l'intermédiaire de l'algorithme *extended advanced Front-End (XAFE)* ETSI ES 202 212 [9, 10]. Cet algorithme a été développé afin de fournir une analyse acoustique robuste au bruit dans le cadre de la reconnaissance de parole distribuée. Le paramètre de pitch est utilisé pour pouvoir traiter le cas des langues tonales et pour permettre la reconstruction du signal de parole.

Dans les expériences décrites par la suite, le pitch est représenté par la valeur de la fréquence fondamentale pour les portions voisées du signal de parole. Pour les trames non voisées, sa valeur est arbitrairement fixée à zéro.

Afin de réduire la dépendance au locuteur, nous utilisons également une valeur normalisée du pitch, $P_{norm}$ :

$$P_{norm}(n) = \frac{P(n)}{\frac{1}{N}\sum_{i=1}^{N} P(i)}$$

où le facteur de dépendance au locuteur est approximée par la moyenne du pitch sur les trames voisées. Nous travaillons également avec le logarithme du pitch et de sa valeur normalisée. La valeur de ces paramètres est fixée arbitrairement à -2 pour les trames non voisées.

### *3.2 Modélisation du Pitch*

Intéressons nous maintenant aux différentes façons d'intégrer le pitch au sein de systèmes de reconnaissance de la parole.

Les HMMs utilisent une estimation de la probabilité d'émission du vecteur acoustique, $x_t$, émis sur l'état $s_n$:

$$p(x_t|s_n) \quad (1)$$

Si on introduit le pitch, la probabilité d'émission devient :

$$p(x_t, y_t|s_n) \quad (2)$$

Où $y_t$ représente le pitch.

Si on considère le pitch comme un coefficient acoustique supplémentaire, l'équation (2) peut s'exprimer sous la forme suivante:

$$p(x_t, y_t|s_n) = p(x_t \,\&\, y_t|s_n) \quad (3)$$

Où *&* signifie que $x_t$ et $y_t$ appartiennent au même vecteur acoustique.

Par ailleurs, en développant l'équation 2, on obtient :

$$p(x_t, y_t|s_n) = p(x_t|y_t, s_n)\, p(y_t|s_n) \quad (4)$$

En considérant que le pitch est indépendant des paramètres acoustiques et qu'il est modélisé de façon gaussienne, l'équation (4) devient :

$$p(x_t, y_t|s_n) = p(x_t|s_n)\, p(y_t|s_n) \quad (5)$$

Où $p(y_t|s_n)$ est une densité de probabilité mono ou multi gaussienne(s).

## 4. RÉSULTATS & DISCUSSION

### *4.1 Pitch dans le vecteur acoustique*

Nous considérons ici les résultats obtenus en introduisant directement au sein du vecteur acoustique le paramètre de pitch pour les différentes normalisations présentées dans la partie 3.1. La probabilité d'émission du vecteur acoustique correspond alors à celle définie dans l'équation (3).

La figure 1 présente les résultats obtenus sur la tâche *Communes*. Il apparaît clairement que l'introduction du pitch améliore les performances de reconnaissance puisque dans tous les cas le *Word Accurracy* est plus élevé que celui obtenu avec le modèle de référence. Il passe, par exemple, de 59,40% avec le modèle de référence à 61,17% en utilisant la valeur brute du pitch, soit une réduction relative de 3% du taux d'erreur. Ceci correspond en fait à une réduction relative de 6% du taux de substitution et de 4% du taux de fausse alarme, pour un taux de rejet à tort inchangé. L'utilisation d'une valeur normalisée du pitch ou d'une représentation logarithmique de celui-ci n'apporte pas de réelles améliorations par rapport à l'utilisation de sa valeur brute.

Les résultats obtenus sur la tâche *Plan Resto* sont aussi représentés sur la figure 1. Contrairement au cas précédent, aucune amélioration n'est obtenue par rapport au modèle de référence, quel que soit le paramètre utilisé (valeur brute ou normalisé, échelle linéaire ou logarithmique). Pour illustrer cette observation on peut remarquer que la meilleure performance est obtenue avec la valeur normalisée du pitch, $P_{norm}$. Le *Word Accurracy* est alors de 72,67 % alors qu'il est de 72.59 % avec le modèle de référence, soit une réduction relative de seulement 0,1%.

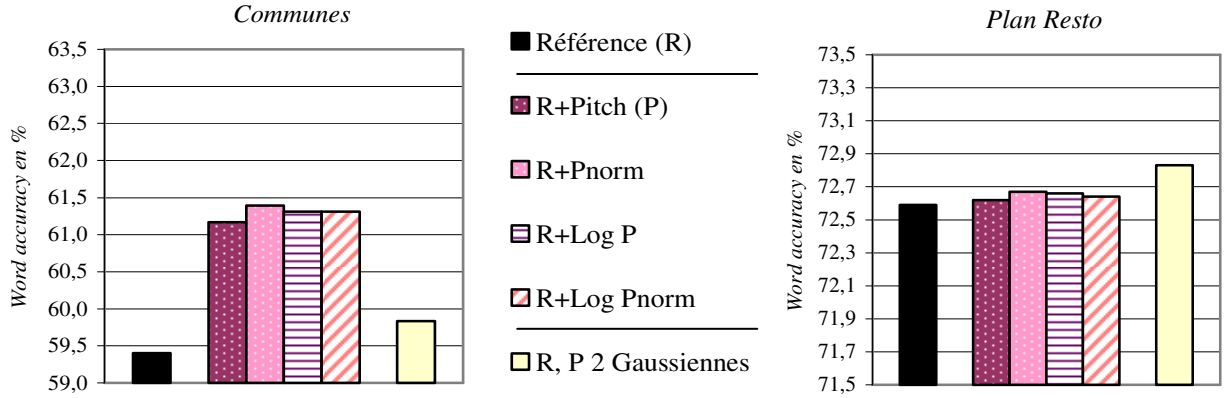


**Figure 1 :** Performances obtenues sur les tâche *Communes* et *Plan Resto*

## *4.2 Corrélation entre pitch et MFCCs*

Les résultats obtenus précédemment montrent clairement une différence de comportement entre les deux tâches étudiées. En effet, l'introduction directe du pitch au sein du vecteur acoustique permet bien d'améliorer les performances de la tâche de reconnaissance de mots isolés, *Communes*. Au contraire ils n'ont quasiment aucune incidence sur la tâche de reconnaissance de parole continue, *Plan Resto*. Ceci est en adéquation avec les différents résultats présentés dans la littérature.

La table 1 représente la matrice de corrélation du logarithme de l'énergie, des coefficients mel-cepstraux et du pitch pour le phonème /œ/ calculée sur les données d'adaptation de la tâche *Communes*. La corrélation entre le pitch et les coefficients acoustiques est représentée dans la dernière colonne et dans la dernière ligne. Les valeurs élevées (supérieures à 0,20) de corrélation sont présentées en gras. On peut observer que pour un certain nombre de MFCCs (C11 et C9 par exemple), la corrélation avec le pitch est bien plus importante qu'avec les autres paramètres acoustiques.

**Table 1 :** Matrice de corrélation entre LogE, MFCCs et pitch calculée sur les données de la tâche *Communes* - Phonème /œ/

| | LogE | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 | Pitch |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LogE | 1.00 | -0.21 | -0.23 | -0.25 | -0.01 | -0.04 | 0.09 | 0.05 | 0.01 | -0.05 | -0.06 | **0.23** |
| C2 | -0.21 | 1.00 | 0.08 | -0.00 | 0.23 | -0.07 | -0.22 | 0.10 | 0.28 | -0.05 | 0.21 | **-0.26** |
| C3 | -0.23 | 0.08 | 1.00 | 0.40 | -0.23 | -0.12 | -0.18 | -0.11 | 0.20 | 0.19 | 0.18 | **-0.34** |
| C4 | -0.25 | -0.00 | 0.40 | 1.00 | -0.25 | -0.24 | -0.04 | -0.10 | 0.03 | 0.09 | 0.28 | **-0.24** |
| C5 | -0.01 | 0.23 | -0.23 | -0.25 | 1.00 | 0.04 | -0.05 | 0.12 | 0.08 | -0.25 | 0.05 | -0.03 |
| C6 | -0.04 | -0.07 | -0.12 | -0.24 | 0.04 | 1.00 | 0.17 | -0.06 | -0.16 | 0.14 | -0.27 | **0.31** |
| C7 | 0.09 | -0.22 | -0.18 | -0.04 | -0.05 | 0.17 | 1.00 | 0.05 | -0.36 | 0.29 | -0.16 | **0.27** |
| C8 | 0.05 | 0.10 | -0.11 | -0.10 | 0.12 | -0.06 | 0.05 | 1.00 | 0.12 | -0.01 | 0.35 | **-0.22** |
| C9 | 0.01 | 0.28 | 0.20 | 0.03 | 0.08 | -0.16 | -0.36 | 0.12 | 1.00 | -0.11 | 0.33 | **-0.56** |
| C10 | -0.05 | -0.05 | 0.19 | 0.09 | -0.25 | 0.14 | 0.29 | -0.01 | -0.11 | 1.00 | -0.07 | -0.02 |
| C11 | -0.06 | 0.21 | 0.18 | 0.28 | 0.05 | -0.27 | -0.16 | 0.35 | 0.33 | -0.07 | 1.00 | **-0.45** |
| Pitch | **0.23** | **-0.26** | **-0.34** | **-0.24** | -0.03 | **0.31** | **0.27** | **-0.22** | **-0.56** | -0.02 | **-0.45** | 1.00 |

La table 2 représente la matrice de corrélation pour le même phonème /œ/ calculée sur les données d'adaptation de la tâche *Plan Resto*. On peut remarquer que le nombre de coefficients de corrélation supérieurs à 0,20 est bien moins important que précédemment : seulement 3 contre 9 pour la tâche *Communes*. Ceci tendrait à montrer que la corrélation entre le pitch et les coefficients cepstraux est moins importante dans le cas de la tâche *Plan Resto*.

Afin de vérifier cette hypothèse, nous avons calculé la corrélation entre le pitch et les différents paramètres acoustiques pour l'ensemble des voyelles communes aux deux tâches. Pour chaque phonème nous avons ensuite relevé le nombre de fois où le coefficient de corrélation entre le pitch et les MFCCs était supérieur à 0,20. Les résultats obtenus sont donnés dans la figure 2 page suivante. On remarque clairement que dans la plupart des cas le nombre de coefficients de corrélation supérieur à 0.20 est plus important pour la tâche *Communes* que pour la tâche *Plan Resto*. Cela signifie que la corrélation entre le pitch et les MFCCs est bien plus grande dans le cas de la tâche de reconnaissance de mots isolés que dans le cas de la tâche de reconnaissance de parole continue. Dans ce dernier cas la prosodie joue un rôle important et peut mener à des variations importantes de pitch selon le contexte prosodique des syllabes. Ceci pourrait expliquer la corrélation moins importante entre le pitch et les MFCCs.

**Table 2 :** Matrice de corrélation entre LogE, MFCCs et pitch calculée sur les données de la tâche *Plan Resto* - Phonème /œ/

| | LogE | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 | Pitch |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LogE | 1.00 | 0.07 | -0.04 | -0.13 | -0.16 | -0.20 | -0.15 | -0.08 | 0.22 | -0.02 | 0.20 | -0.00 |
| C2 | 0.07 | 1.00 | -0.09 | -0.27 | 0.32 | 0.04 | -0.14 | -0.08 | 0.25 | -0.10 | 0.07 | -0.13 |
| C3 | -0.04 | -0.09 | 1.00 | 0.44 | -0.19 | -0.10 | -0.03 | -0.28 | 0.00 | 0.05 | 0.08 | -0.10 |
| C4 | -0.13 | -0.27 | 0.44 | 1.00 | -0.33 | -0.20 | 0.29 | -0.16 | -0.27 | 0.08 | 0.26 | -0.07 |
| C5 | -0.16 | 0.32 | -0.19 | -0.33 | 1.00 | 0.12 | -0.23 | 0.04 | 0.24 | -0.32 | -0.04 | -0.04 |
| C6 | -0.20 | 0.04 | -0.10 | -0.20 | 0.12 | 1.00 | -0.09 | -0.07 | 0.04 | 0.15 | -0.29 | 0.19 |
| C7 | -0.15 | -0.14 | -0.03 | 0.29 | -0.23 | -0.09 | 1.00 | 0.12 | -0.40 | 0.35 | 0.09 | **0.21** |
| C8 | -0.08 | -0.08 | -0.28 | -0.16 | 0.04 | -0.07 | 0.12 | 1.00 | -0.05 | 0.01 | 0.15 | 0.11 |
| C9 | 0.22 | 0.25 | 0.00 | -0.27 | 0.24 | 0.04 | -0.40 | -0.05 | 1.00 | -0.37 | 0.03 | **-0.39** |
| C10 | -0.02 | -0.10 | 0.05 | 0.08 | -0.32 | 0.15 | 0.35 | 0.01 | -0.37 | 1.00 | -0.13 | **0.24** |
| C11 | 0.20 | 0.07 | 0.08 | 0.26 | -0.04 | -0.29 | 0.09 | 0.15 | 0.03 | -0.13 | 1.00 | -0.19 |
| Pitch | -0.00 | -0.13 | -0.10 | -0.07 | -0.04 | 0.19 | **0.21** | 0.11 | **-0.39** | **0.24** | -0.19 | 1.00 |

## *4.3 Modélisation indépendante du Pitch*

Nous avons également modélisé le pitch indépendamment des autres paramètres acoustiques par des distributions à 1, 2, 4 et 8 gaussiennes. La probabilité d'émission correspond cette fois à celle définie dans l'équation (5). Les meilleures performances furent obtenues avec une modélisation à 2 gaussiennes. Les résultats obtenus sur les deux tâches avec cette modélisation sont présentés sur la figure 1. Ici une légère amélioration des performances est observée dans les deux cas. Notons toutefois que pour

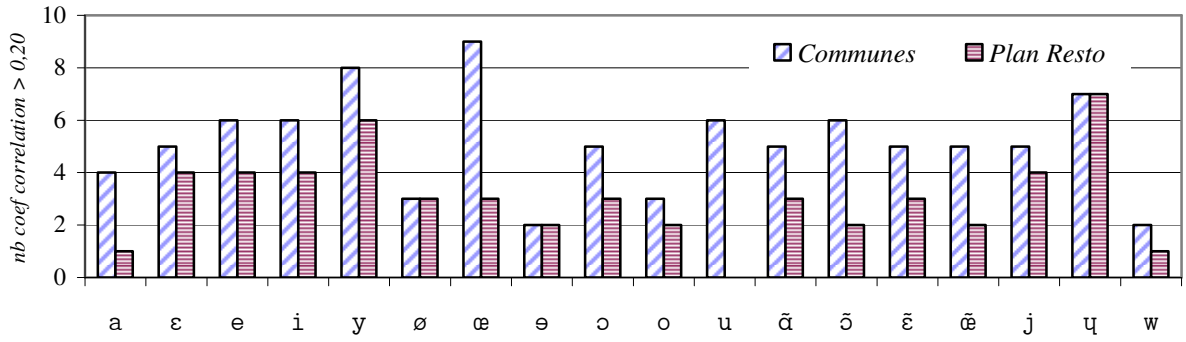


**Figure 2 :** Comparaison de la corrélation entre le pitch et les coefficients acoustiques pour les tâches *Communes* et *Plan Resto*

la tâche *Communes* cette amélioration est loin d'atteindre celle obtenue précédemment (Le *Word Accuracy* obtenu est de 59,84 % alors qu'il est de 61,17% lorsque le pitch est directement intégré au vecteur acoustique).

Ceci montre néanmoins que les deux tâches de reconnaissance de mots isolés et de parole continue se comportent de façon similaire vis-à-vis du pitch lorsqu'on ne peut tirer profit d'une certaine corrélation entre le pitch et les MFCCs. Ceci montre donc que c'est bien la corrélation plus ou moins importante entre le pitch et les autres paramètres acoustiques qui peut expliquer les différences de résultats entre les deux tâches *Communes* et *Plan Resto* obtenues précédemment en intégrant directement le paramètre de pitch au sein du vecteur acoustique.

## 5. CONCLUSION

Dans ce papier nous avons comparé le rôle joué par le pitch sur les performances de deux tâches de reconnaissance indépendante du locuteur. En introduisant directement le pitch au sein du vecteur acoustique, les performances de la tâche de reconnaissance de mots isolés ont été nettement améliorées, alors que celles de la tâche de reconnaissance de parole continue sont restées quasiment inchangées. Au contraire en modélisant le pitch indépendamment des autres coefficients acoustiques, nous avons obtenu un comportement similaire des deux tâches de reconnaissance. Ces résultats pourraient s'expliquer par le fait que l'amélioration apportée en intégrant directement le pitch au vecteur acoustique soit liée à la corrélation entre le pitch et les coefficients cepstraux. Cette corrélation est plus importante dans le cas de la tâche *Communes*, qui est une tâche de reconnaissance de mots isolés, que dans le cas de la tâche *Plan Resto*, tâche de reconnaissance de parole continue dans laquelle la prosodie peut mener à des variations importantes du pitch selon le contexte prosodique.

## BIBLIOGAPHIE


[1] A. Martin & L. Mauuary, "Voicing Parameter and Energy-Based Speech/Non-Speech Detection for Speech Recognition in Adverse Conditions", in *Proc. Eurospeech'2003, European Conf. on Speech Communication and Technology*, Genève, Suisse, 1-4 Sept. 2003.

[2] S. Liu, S. Doyle, A. Morris & F. Ehsam, "The effect of fundamental frequency on Mandarin speech recognition", in *Proc. ICSLP'98, Int. Conf. on Spoken Language Processing*, Sydney, Australie, vol. 6, pp 2647-2650, 30 Nov.–4 Dec. 1998.

[3] A Ljolje, "Speech Recognition Using Fundamental Frequency and Voicing in Acoustic Modeling", in *Proc ICSLP'2002*, *Int. Conf. on Spoken Language Processing*, Denver, USA, pp 2137-2140, 16-20 Sept. 2002.

[4] M. Graciarena, H. Franco, J. Zeng, D. Vergyri, A. Stolke, "Voicing feature integration in SRI's decipher LVCSR System", in *Proc. ICASSP'2004*, *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Montreal, Canada, vol. 1, pp 921-924, 17-21 May 2004.

[5] G.Cloarec, J.Monné & D.Jouvet, "Introducing Pitch and Voicing Parameters into Speaker-Independent Speech Recognition Systems", in *Proc. SPECOM'2005*, *10th Int. Conf. on Speech and Computer*, Patras, Grèce, pp95-98, 17-19 Oct. 2005.

[6] Denis Jouvet, K. Bartkova, L. Delphin-Poulat, A. Ferrieux, X. Lamming, J. Monné & C. Raix, "About improving recognition of spontaneously uttered French city names", in *Proc. ICASSP'2003, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Hong-Kong, vol. 1, pp 544-547, April 2003.

[7] M.D Sadek, A. Ferrieux, A. Ozannet, P. Bretier, F. Panaget, J. Simonin, "Effective Human-Computer cooperative spoken dialogue: the AGS demonstrator", in *Proc. ICSLP'96*, *Int. Conf. on Spoken Language Processing*, Philadelphie, USA, vol. 1, pp 546-549, 3-6 Oct. 1996

[8] C. Raymond, F. Béchet, N. Camelin, R. de Mori, G. Damnati, "Semantic interpretation with error correction", in *Proc. ICASSP'2005, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphie, USA, vol. 1, pp 29-32, March 2005.

[9] ETSI ES 202 212 V1.1.1 (STQ); *Distributed speech recognition; Extended advanced front-end feature extraction*.

[10] A. Sorin, T. Ramabadran, D. Chazan, R. Hoory, M. McLaughlin, D. Pearce, F. CR Wang & Y. Zhang, "The ETSI Extended Distributed Speech Recognition (DSR) Standards: client side processing and tonal language recognition evaluation", in *Proc. ICASSP'2004, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Montreal, Canada, vol. 1, pp 53-56, 17-21 May 2004.