# Amélioration de la reconnaissance de la parole issue de téléphones GSM sans données réelles

*Jérôme Baude, Nancy Bertin, Claire Waast-Richard*

IBM France - EMEA Voice Technology Development
Tour Descartes, 2 avenue Gambetta - 92066 Paris-La Défense
tél. : +33 1 49 05 78 88
mél. : jerome.baude@fr.ibm.com, nbertin@enst.fr, waast@fr.ibm.com

## ABSTRACT

This paper investigates the influence of GSM speech coding on speaker-independent speech recognition, and suggests a possible way to improve speech recognition performances when speech passes through a mobile phone network using GSM standards. We first evaluated the performance degradation caused by GSM transcoding. Then, we tried to improve the performances by modelling GSM transcoding during the training phase. Since real mobile data are not readily available, especially French ones, we simulated such data by transcoding artificially a set of landline phone data. By training an acoustic model with this simulated data set, we managed to improve significantly the recognition performances on real mobile phone data.

## 1. INTRODUCTION

L'extraordinaire expansion que connaît aujourd'hui la téléphonie mobile pose de nouveaux défis aux technologies vocales, notamment à la reconnaissance de la parole. La téléphonie mobile s'accompagne de contraintes (bruit ambiant, fort taux de compression...) risquant de perturber les performances des systèmes de reconnaissance de la parole. Le développement d'applications vocales attractives passe nécessairement par leur efficacité sur téléphone portable ; améliorer les performances des systèmes utilisés dans ces conditions est donc une forte préoccupation dans le monde des technologies vocales.

Les systèmes actuels de reconnaissance de la parole sont basés sur des méthodes d'apprentissage statistique. La modélisation statistique d'un phénomène physique est d'autant plus fine que ce phénomène est fortement présent dans les données d'apprentissage. Or, les données issues de téléphonie mobile et exploitables pour la construction de tels modèles sont à la fois peu nombreuses (notamment en langue française) et coûteuses à acquérir.

Dans cet article, nous nous sommes plus précisément intéressés à l'impact du codage sur les performances en reconnaissance de la parole indépendante du locuteur, et sur une méthode de compensation de cet impact par simulation. En effet le codage de la parole réalisé avant transmission sur le réseau est une compression avec perte, dégradant le signal. Nos travaux ont porté sur les deux versions les plus récentes des codecs GSM (Global System for Mobile communications, le standard européen de téléphonie mobile) déployées en France, le Half-Rate et le Enhanced-Full-Rate, dont le fonctionnement est brièvement décrit dans la *section* 2. Après avoir évalué la dégradation dûe au codage (voir *section* 3) nous avons expérimenté l'utilisation de données simulées (artificiellement transcodées) à l'apprentissage et son effet sur les performances en reconnaissance (*section* 4). Enfin, la *section* 5 conclut ces travaux et examine de possibles suites à cette étude.

## 2. CODECS GSM

Le GSM (Global System for Mobile communications) est le standard pour la téléphonie mobile dans toute l'Europe. La norme propose trois algorithmes de codage de la parole désignés par "full rate", "half rate" et "enhanced full rate". Seuls les deux derniers ont été utilisés ici. Le rôle des codecs GSM est de compresser le signal de parole avant sa transmission, de manière à réduire le nombre de bits nécessaires à sa représentation, tout en maintenant une qualité de signal acceptable en sortie. Il s'agit dans tous les cas d'une compression avec perte, affectant les propriétés spectrales du signal. Les trois versions du codec prennent en entrée des fichiers audio-numériques au format PCM, échantillonnés à 8 kHz et représentés sur 13 bits. Chaque trame en entrée, d'une durée de 20 ms, est traitée séparément. Entre les trois codeurs, c'est principalement l'algorithme de quantification vectorielle qui varie.

### *2.1. Le codeur "Half-Rate" HR*

Le codeur "Half-Rate", fonctionnellement décrit dans [7], fonctionne au débit de 5,6 kbits/s, atteint grâce au principe de codage dit "VSELP" (*Vector Sum Excited Linear Prediction*) proposé dans [3]. D'autre part, les sources C-ANSI d'un programme réalisant le codage-décodage selon cette norme sont disponibles en [5]. Nous avons utilisé ce programme pour nos simulations.

### *2.2. Le codeur "Enhanced Full Rate" EFR*

Le codeur "Enhanced Full-Rate" est le plus récent des trois codeurs définis par la norme. Sa description fonctionnelle est disponible dans [6] et, comme précédemment, [4] fournit le code C-ANSI mettant en oeuvre la norme. Il est construit autour de l'algorithme ACELP (Algebraic Code Excited Linear Prediction) et fonctionne avec un débit de sortie de 12,2 kbits/s. Il est réputé pour fournir un signal de sortie de qualité auditive nettement supérieure aux deux autres codecs.

### *2.3. Quantification*

Contrairement au codec utilisé en situation réelle, qui prend en entrée un signal numérisé directement sur 13 bits

linéaires, le codec que nous avons utilisé traite des fichiers 16 bits. Le programme se charge lui-même du passage sur 13 bits par une opération de troncature : les trois bits de poids le plus faible sont forcés à zéro. Cette opération correspond à une sous-quantification et donc à une perte d'information qui sera d'autant plus importante que les extraits de parole n'auront pas été numérisés à pleine échelle lors de leur acquisition. Ce problème a notamment été soulevé dans [2]. Lorsque cela était possible, nous avons effectué une préamplification uniforme du signal par simple décalage binaire vers les poids forts, de façon à sauvegarder au maximum les bits risquant d'être tronqués par le programme.

### *2.4. Détection d'activité vocale et option de transmission discontinue*

En vue de réduire la consommation en énergie des téléphones et soulager l'encombrement des réseaux, la norme GSM prévoit la possibilité d'interrompre la transmission de données lors des périodes de silence. Deux modules gèrent cette fonctionnalité : le module dit "VAD" (*Voice Activity Detection*, i.e. détection d'activité vocale) décide de la présence ou de l'absence de parole ; le module "DTX" (*Discontinuous Transmission* ou transmission discontinue) gère l'activation ou l'arrêt de la transmission (en fonction de l'information fournie par le VAD). Lors des périodes de non transmission, le récepteur ajoute un bruit dit "de confort" pour éviter un silence désagréable à l'interlocuteur (impression d'avoir perdu la communication). Cette fonctionnalité étant optionnelle, nous avons étudié son influence sur nos données au décodage et à l'apprentissage. En particulier, l'article [2] fait référence à une dégradation supplémentaire introduite par l'activation de cette option.

## 3. INFLUENCE DU CODEC SUR LES PERFORMANCES DE DÉCODAGE

Les causes possibles de dégradation de performances dûes à l'utilisation d'un téléphone portable sont nombreuses : le bruit, la perte de paquets, l'influence du canal... [9]. En particulier, nous avons pu consulter dans la littérature plusieurs articles démontrant l'impact du codage de la parole sur les performances de systèmes vocaux, notamment en ce qui concerne la reconnaissance du locuteur à partir de parole codée suivant la norme GSM (voir [2] et [10]) mais aussi la reconnaissance de la parole codée suivant différents standards (par exemple dans [1] et [8]). Nous avons souhaité dans un premier temps vérifier et évaluer cette dégradation dans le cadre plus précis de nos travaux (codeurs GSM half-rate et enhanced-full-rate, tâche de reconnaissance de la parole indépendante du locuteur).

Notre première expérience a consisté à comparer les performances d'un système de référence au décodage de phrases originales (enregistrées sur téléphone fixe) et de leurs versions transcodée (codée-décodée) par un codec, afin d'évaluer quantitativement l'impact du codage sur la reconnaissance de la parole.

### *3.1. Bases de test et d'apprentissage*

Nous avons d'abord sélectionné un jeu de test de 5.000 phrases enregistrées sur téléphone fixe, couvrant une dizaine de locuteurs différents et absents de tous les corpus d'apprentissage, et une variété suffisante de tâches : validation (oui/non), paires Prénoms-Noms, mots de commandes simples (répéter, suivant, précédent...), dates, numéros de téléphones... Cette base (T-Ref) a ensuite subi les traitements correspondants aux codecs que nous souhaitions caractériser :

- Un codage Enhanced-Full-Rate avec les options par défaut, c'est-à-dire avec activation de l'option de détection d'activité vocale et transmission discontinue, sans traitement supplémentaire (T-EFR-DTX).
- Un codage Enhanced-Full-Rate sans cette option de transmission discontinue (T-EFR-noDTX).
- et enfin le codage Half-Rate, dans les mêmes conditions (pas de détection d'activité vocale) (T-HR-noDTX).

La séquence de pré-traitement a consisté en une application successive du codeur puis du décodeur. Les données obtenues sont des "simulations" de données mobiles, isolant particulièrement l'opération de codage-décodage propre à ces données.

D'autre part, on a construit un modèle acoustique de référence avec des données exclusivement issues de téléphonie fixe, qui est utilisé pour le décodage (M-Ref).

### *3.2. Méthodologie*

Etant donné le modèle acoustique de référence, nous avons réalisé les décodages du test fixe initial et de ses versions transcodées et avons comparé les résultats, entre lesquels seuls les codages appliqués à la parole diffèrent, ce qui permet d'isoler l'influence propre du codec. Le critère d'évaluation retenu pour cette comparaison est le taux d'erreurs par mot.

### *3.3. Résultats*

Le tableau 1 présente les résultats de décodage de la base de test initiale et de ses versions transcodées half-rate et enhanced-full-rate. De manière prévisible, on observe une dégradation significative des performances en reconnaissance. Si l'on exprime cette dégradation relativement entre le premier test (décodage des données initiales) et les tests transcodés) on obtient une dégradation relative de 15 % à 40 %. Le pire cas correspond très logiquement au codec Half-Rate qui dégrade davantage le signal de parole.

**TAB. 1:** Résultats de décodage des bases de test initiale et transcodées par le modèle acoustique de référence, exprimés en Taux d'erreurs par mot

| | **M-ref** |
|---|---|
| **T-ref** | 5,8 % |
| **T-EFR-noDTX** | 6,5 % |
| **T-EFR-DTX** | 6,9 % |
| **T-HR-noDTX** | 9,7 % |

Par ailleurs, on observe ici la dégradation engendrée par l'activation de l'option de transmission discontinue "DTX".

### *3.4. Commentaires*

Ces résultats confirment les conséquences néfastes du codage-décodage de la parole sur les performances en reconnaissance de parole indépendante du locuteur, même si la dégradation générale des performances lors de l'utilisation d'un téléphone portable n'est pas entièrement imputable à ce transcodage.

La version la plus ancienne des deux codecs testés, le codec Half-Rate, provoque à la fois la plus forte dégradation de la qualité auditive du signal (en raison de son fort taux de compression) et la plus forte diminution de performance du système de reconnaissance utilisé.

Nous nous sommes demandés si cette déformation du signal était une perte sèche d'information (causant une irrémédiable dégradation des performances du système) ou un phénomène modélisable par l'apprentissage de données ayant subi un traitement similaire.

## 4. MODÉLISATION DU CODEC EN PHASE D'APPRENTISSAGE

On cherche à confirmer l'espoir de modélisation des effets du transcodage GSM lors de l'apprentissage. L'idée sous-jacente est de pouvoir, de manière économique, ajouter aux corpus d'apprentissage des phrases ayant subi ce transcodage. L'idéal serait de pouvoir utiliser des données réellement issues de téléphonie mobile mais elles sont rares et coûteuses à acquérir ; nous avons cherché à savoir si simuler de telles données en reproduisant artificiellement la séquence codage-décodage pouvait compenser partiellement la perte en situation réelle.

On construit tout d'abord des modèles acoustiques comparables au modèle de référence utilisé dans la *section* 3 et ne différant de celui-ci que par le pré-traitement de codage-décodage appliqué. On préamplifie le signal de façon à réduire la perte sèche due à la quantification mentionnée *section* 2.3.

La totalité des données du modèle de référence subit une séquence codage-décodage.

Au final, quatre modèles sont donc calculés : le modèle de référence (M-Ref) et ses trois versions, correspondant aux trois codages-décodages applicables : enhanced full-rate avec DTX (M-EFR-DTX) enhanced full-rate sans DTX (M-EFR-DTX), half-rate sans DTX (M-HR-noDTX).

### *4.1. Vérification de la modélisation du codec*

Le premier objectif est de s'assurer que les modèles construits modélisent les effets du codage. Pour vérifier cela, on cherche à s'assurer que pour un traitement donné, le décodage d'un test transcodé avec un modèle ayant subi le même traitement donne un meilleur score que le décodage de ce même test avec le modèle de référence.

On observe dans la table 2 un gain pour chacun des trois modèles. Cependant, les effets de la transmission discontinue semblent peu modélisables. Le modèle Enhanced Full Rate avec DTX (M-EFR-DTX) ne présente qu'une amélioration de 9% alors que les modèles sans DTX (M-EFR-noDTX et M-HR-noDTX) présentent, respectivement, une amélioration de 17 et 25%.

**TAB. 2:** Comparaison croisée du modèle de référence (M-Ref) et ses versions transcodées sur les tests transcodés (M-EFR-noDTX, M-EFR-DTX, M-HR-noDTX)

| | **M-ref** | **M-EFR noDTX** | **M-EFR DTX** | **M-HR noDTX** |
|---|---|---|---|---|
| **T-EFR noDTX** | 6,5 % | 5,4% | - | - |
| **T-EFR DTX** | 6,9 % | - | 6,3% | - |
| **T-HR noDTX** | 9,7 % | - | - | 7,3 % |

Ces résultats montrent qu'une modélisation des effets du codage est possible. Il reste maintenant à valider cette modélisation en analysant les performances de ces modèles sur un jeu de test réel.

### *4.2. Validation de la modélisation sur un jeu de test réel (T-réel)*

Le jeu de test utilisé est constitué de 1.200 phrases. Ces données ont été enregistrées, depuis différents téléphones portables, par un nombre important de locuteurs. Elles couvrent 3 tâches : commandes, validation (oui/non), noms communs. L'enregistrement est effectué dans la rue par des locuteurs naïfs. L'élocution n'est pas toujours parfaite. Un fort bruit ambiant est présent ainsi que la superposition de paroles provenant de tierces personnes. A bien des égards, ce test est très difficile. Ces données ayant été enregistrées récemment, le codec utilisé est donc l'EFR. Les résultats de cette expérience sont résumés dans le tableau 3.

**TAB. 3:** Résultats du décodage du test de référence (T-ref) et du test mobile réel (T-réel) pour le modèle de référence (M-ref) et les modèles transcodés (M-EFR-noDTX, M-EFR-DTX, M-HR-noDTX)

| | **M-ref** | **M-EFR noDTX** | **M-EFR DTX** | **M-HR noDTX** |
|---|---|---|---|---|
| **T-ref** | **5,8%** | **5,9%** | 6,5% | 6,1% |
| **T-réel** | **16,2%** | **15,2%** | 16,7% | 16,6% |

Le modèle ayant subi un transcodage Half-Rate est moins bon que le modèle de référence. Les caractéristiques EFR des données de test ne sont pas modélisées par un transcodage Half-Rate.

Le modèle ayant subi un transcodage Enhanced-Full-Rate avec détection d'activité vocale est aussi moins bon que le modèle de référence. La transmission discontinue du signal est une perte d'information sèche ne pouvant être modélisée à l'apprentissage.

Le modèle ayant subi un transcodage Enhanced-Full-Rate sans détection d'activité vocale présente une amélioration de 6% par rapport au même modèle construit à partir des données originales.

Cette amélioration observée sur le test mobile réel est associée d'une dégradation sur le jeu de test fixe. Celle-ci

reste très faible puisqu'on passe de 5,8% à 5,9% pour le modèle Enhanced Full Rate sans DTX (M-EFR-noDTX).

On est donc parvenu à modéliser l'opération de codage-décodage lors de la phase d'apprentissage sans trop dégrader les performances sur le jeu de test fixe T-ref. L'utilisation de données simulées permet de pallier, en partie, l'absence de données réellement issues de téléphones mobiles.

Il serait intéressant de vérifier si cette amélioration reste toujours valable en intégrant des données mobiles réelles à l'apprentissage. De même, nous pourrions doubler les phrases issues de la téléphonie fixe en intégrant dans le corpus d'apprentissage la version originale et la version transcodée afin de voir si on peut obtenir une amélioration sur le jeu de test mobile sans dégrader les performances sur le jeu de test fixe.

## 5. CONCLUSION

Au cours de cette étude nous avons pu simuler le codage-décodage GSM sur des données issues de la téléphonie fixe, observer son impact sur un système de reconnaissance de la parole et modéliser ses caractéristiques lors de l'apprentissage d'un modèle acoustique en français.

La dégradation due au codec concourt de manière importante à la dégradation globale observée lors du décodage de données mobiles. Cette dégradation a pu être en partie compensée en introduisant dans les données d'apprentissage du modèle, des données issues de téléphonie fixe ayant subi un codage-décodage GSM artificiel. Cette simulation permet de remplacer avantageusement de vraies données mobiles, difficiles à obtenir, lors de la construction de modèles pour les systèmes de reconnaissance de la parole destinés à être contactés depuis un téléphone portable.

## 6. REMERCIEMENTS

Les auteurs remercient Rémi Lejeune et Catherine Tchong pour leur aide régulière sur ces expériences. Rémi a plus particulièrement participé à la construction du modèle de référence tandis que Catherine nous a aidé à résoudre certains "bugs" rencontrés lors de la construction des modèles transcodés.

Enfin nous remercions l'ensemble de l'équipe pour leur participation aux discussions, et pour le travail quotidien sans quoi rien ne serait faisable.

## RÉFÉRENCES

[1] L. Besacier, C. Bergamini, D. Vaufreydaz, and E. Castelli. The effect of speech and audio compression on speech recognition performance. In *IEEE Multimedia Signal Processing Workshop*, Cannes, France, October 2001.

[2] L. Besacier, S. Grassi, A. Dufaux, M. Ansorge, and F. Pellandini. GSM speech coding and speaker recognition. In *ICASSP 2000*, Istanbul, Turkey, 5–9 June 2000.

[3] I. Gerson and M. Jasiuk. A 56oo bps VSELP speech coder candidate for half rate GSM. In *Proc. Eurospeech'93*, volume 1, pages 253–256, 1993.

[4] European Telecommunications Standard Institute. Digital cellular telecommuncications system (phase 2+), ANSI-C code for the GSM enhanced full rate (EFR) speech codec ; GSM 6.53. In *ETSI EN 300 724 v8.0.1*, 1999.

[5] European Telecommunications Standard Institute. Digital cellular telecommuncications system (phase 2+), ANSI-C code for the GSM half rate (HR) speech codec ; GSM 6.06. In *ETSI EN 300 969 v8.0.1*, 1999.

[6] European Telecommunications Standard Institute. Digital cellular telecommunications system (phase 2+), enhanced full rate (EFR) speech transcoding ; GSM 6.60. In *ETSI EN 300 726 v8.0.1*, 1999.

[7] European Telecommunications Standard Institute. Digital cellular telecommunications system (phase 2+), half rate (HR) speech transcoding ; GSM 6.20. In *ETSI EN 300 967 v8.0.1*, 1999.

[8] B.T. Lilly and K.K. Paliwal. Effect of speech coders on speech recognition performance.

[9] S. Möller and H. Bourlard. Analytic assessment of telephone transmission impact on asr performance using a simulation model. *Speech communication*, 38 :441–459, 2002.

[10] M. Phythian, J. Ingram, and S. Sridharan. Effects of speech coding on text-dependent speaker recognition. In *IEEE TENCON speech and image technologies for computing and telecommunications*, 1997.

[11] T. Salonidis and V. Digilakis. Robust speech recognition for multiple topological scenarios of the GSM mobile phone system. 1998.