Evaluation de la Reconnaissance de la Parole VoIP avec Dissimulation de Perte de Paquets

Adil Bakri, Abderrahmane Amrouche

الملخّـص

نقترح في هذه الدّراسة استعمال تقنية إخفاء IP (VoIP) لزيادة قوّة التّعرّف الآلي على الكلام، على الشّبكة. هذه التقنية تقوم على إنشاء إشارة كلاميّة مركّبة من خلال انتقال مرن بين إشارة حقيقيّة الحزم الضائعة (PLC) مع G711 ApendixI ذات نوع (PLC) إلى إشارة اصطناعيّة. وفي الجزء المكرّس للتعرّف الآليّ على الكلام قمنا بتكييف تقنية HTK، مطبّقين نظاما مفتوح المصدر G729. المرمز يتم محاكاة الحزمة الضائعة بنموذج ماركوف مع حالتين. ظهر تحسن كبير في معدل IP المستخدم في شبكات الاتصالات.

الكلمات المفاتيح : الكلام على الإنترنت، بروتوكول (G729)، مرمز (OLA)، تغطية الإضافة (RAP)، التّعرّف الآلي على الكلام (VoIP).

Evaluation de la Reconnaissance de la Parole VoIP avec Dissimulation de Perte de Paquets

Adil Bakri[1], Abderrahmane Amrouche[2]

[1]Centre de Recherche Scientifique et Technique pour le Développement de Langue Arabe (CRSTDLA), Alger, Algérie.
[2]Laboratoire de Communication parlée et de Traitement des Signaux Faculté d'Electronique et d'Informatique, (USTHB), Alger Algérie.
adil.msilib@yahoo.fr, namrouche@usthb.dz

Résumé

Pour augmenter la robustesse de la Reconnaissance Automatique de la Parole (RAP) sur le réseau IP (VoIP), nous proposons dans cet article l'utilisation du masquage de perte de paquets (PLC : Packet Loss Concealment). Cette méthode consiste à générer un signal vocal synthétique destiné à remplacer les données manquantes, en assurant une transition douce entre le signal réel et le signal synthétique. Ainsi, dans ce travail nous avons adapté la recommandation ITU-I G711 Appendix I au codec G729. Pour la partie reconnaissance de parole, nous avons implémenté le système open source HTK (Hidden Markov Models ToolKit), alors que la perte de paquets est simulée par un modèle de Markov à deux états. Les résultats expérimentaux avec de la parole transcodée avec le codec G729 utilisé dans les réseaux VoIP montrent une amélioration sensible du taux de reconnaissance avec la méthode de masquage des pertes de paquets développée, confortant ainsi la démarche suivie dans ce travail.

Mots Clés : VoIP, RAP, OLA, PLC, G729, ITU-I G711 Appendix I, HMM.

1. Introduction

Dans un système VoIP, au niveau du récepteur, certains paquets peuvent manquer, à cause des délais, à l'encombrement, ou aux erreurs de transfert. Dans les réseaux de communications, ces pertes sont causées par plusieurs facteurs liés aux différentes étapes de la chaîne de transmission en VoIP, en particulier la congestion des nœuds (routeurs). Nous savons par ailleurs que la perte des paquets cause une perte de synchronisation entre le codeur et le décodeur.

La perte de paquets dégrade la qualité de la voix et influe sur la qualité de la parole. Elle se traduit par des ruptures au niveau de la conversation et une impression de hachure de la parole. Il est, par conséquent, indispensable de mettre en place un mécanisme de dissimulation de perte de paquets. Plusieurs algorithmes de masquage des pertes de paquets PLC sont utilisés, aussi bien au niveau de l'émetteur qu'au niveau du récepteur.

Dans ce travail, nous nous intéressons à l'étude de l'effet des pertes de paquets sur les performances des systèmes de Reconnaissance Automatique de la Parole (RAP). A ce titre nous avons implémenté la technique de dissimulation de perte de

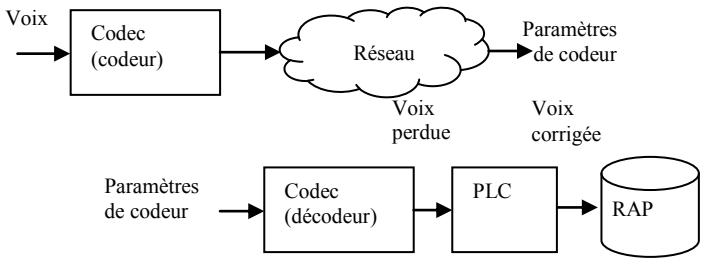


Figure 1: Transmission de voix sur réseau IP.

paquets PLC basée sur ITU-I G711 Appendix I. Le système de RAP mis en œuvre a été réalisé en utilisant la plateforme open source HTK (Hidden Markov Model ToolKit) [1]. Nous avons utilisé la base de données ARADIGIT8K qui a été passée à travers le Codec G.729 pour obtenir la base de données transcodée G.729, à savoir la base ARADIGIT_G729. Cet article est organisé comme suit : après une brève introduction, nous décrirons dans la section 2 le principe de la transmission dans les réseaux VoIP, puis les techniques de masquage des pertes de paquets. La reconnaissance automatique de parole fera l'objet de la troisième section alors que les résultats expérimentaux seront présentés dans la section 4. Nous terminons cet article par une conclusion et les travaux futurs.

2. Technique de dissimulation de la perte de paquets

2.1. Transmission de la Parole en VoIP

Les réseaux de transmission VoIP utilisent des codecs, principalement le G711. Mais en raison de son débit élevé (64Kbits/s), il commence à être supplanté progressivement par le G.729 de débit nettement inférieur. Le codec de parole G.729 est basé sur l'algorithme de prédiction CS-ACELP (Prédiction Linéaire avec Excitation par séquence Codés à Structure Algébrique Conjuguée) et opère sur des trames de parole de 10 ms qui correspondent à 80 échantillons numérisés sur 16 bits pour une fréquence d'échantillonnage de 8 kHz [2]. Le signal de parole est analysé pour extraire les paramètres de codeur et envoyé par paquets à travers le réseau IP [3]. Le décodeur utilise ces paramètres pour reconstruire un signal de parole synthétique (Figure .1) [4].

2.2. Modèle du réseau

Nous avons employé un modèle simple de réseau appelé modèle de Markov à deux états pour modéliser le processus point à point de pertes des paquets sur le réseau IP. L'état 0 indique que le paquet est reçu et l'état 1 qu'il est perdu. La figure 2 représente les pertes de paquets modélisée par un processus aléatoire de Markov à deux états.

Soit p la probabilité pour que le modèle du réseau abandonne un paquet sachant que le paquet précédent est livré, c'est à dire la probabilité de transiter de l'état 0 à l'état 1.

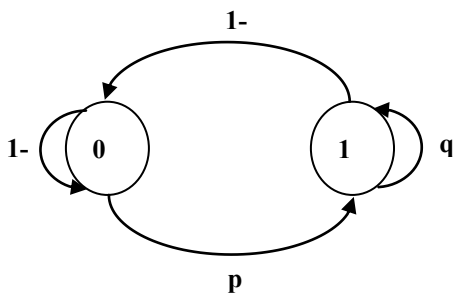


Figure 2: Packet loss modeled by a Markov random process.

Soit q la probabilité pour que le modèle du réseau abandonne un paquet sachant que le paquet précédent est abandonné, c'est à dire la probabilité pour que le modèle reste dans l'état 1. Cette probabilité est également connue comme la probabilité conditionnelle de perte. Soient les probabilités pour rester dans l'état 0 et l'état 1 respectivement :

$$P_0 = \frac{1-q}{p+1-q}$$

$$P_1 = \frac{p}{p+1-q}$$

2.3. Masquage de perte de paquets basé sur la reconnaissance ITU-T G711 Appendix I .

Le masquage PLC basé sur ITU-T G711 Appendix I consiste à générer un signal vocal synthétique destiné à remplacer les données manquantes (effacées) d'un train binaire comme le montre la figure 3. Ce signal synthétique aura de préférence le même timbre et les mêmes caractéristiques spectrales que le signal manquant. Etant donné que les signaux vocaux ont souvent une redondance [5], il est possible de se référer à l'historique des signaux précédents pour produire un son synthétique raisonnablement proche de celui correspondant au segment manquant [6].

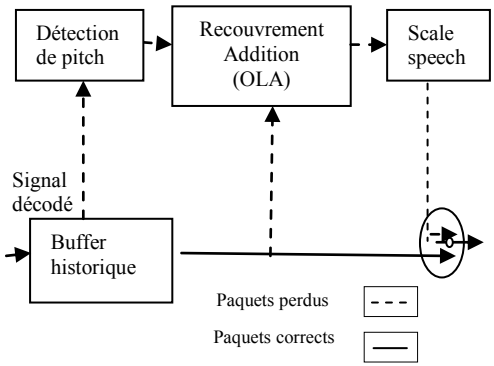


Figure 3: Architecture du système de masquage de perte par technique ITU-T G711 Appendix I.

2.3.1. Buffer historique

La technique PLC enregistre une copie du résultat décodé dans un buffer historique de 48,75 ms (390 échantillons) comme indiqué dans la figure 4. Ce tampon sert à calculer la période fondamentale du moment et à extraire des signaux effacés. Cette mise en tampon n'introduit aucun retard dans le signal de sortie [6].

2.3.2. Détection de Pitch

La période fondamentale est calculée en déterminant la crête de l'intercorrélation normalisée entre les 20 ms les plus récentes des sons vocaux dans le tampon historique et les sons vocaux précédents au moyen de prélèvements de 5 (40 échantillons) à 15 ms (120 échantillons) de parole.

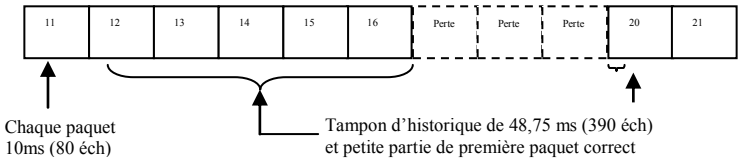


Figure 4 : Exemple avec trois paquets perdus et buffer historique.

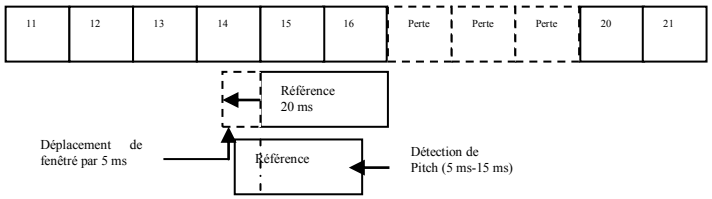
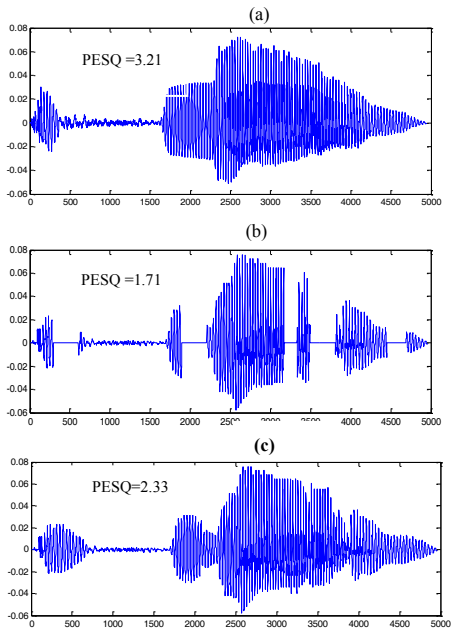


Figure 5: Fenêtre de corrélation pour la détection du Pitch.

La Figure 5 représente la dernière fenêtre de 20 ms de parole précédent l'effacement, il s'agit du signal de référence utilisée pour la détection du pitch au moyen de la corrélation croisée normalisée. Cette fenêtre recule par étapes de 40 à 120 échantillons [7].

2.3.3. Technique de recouvrement addition

La technique de recouvrement addition OLA (Over Lap and Add), assure une transition douce entre les signaux réels et les signaux synthétiques. Pour créer le tampon tonal, le quart de longueur d'onde (la position de période fondamentale dans le tampon tonale) qui précède l'effacement subit une opération OLA au moyen d'une fenêtre triangulaire au quart de longueur d'onde de la période fondamentale précédente. La fenêtre de synthèse est appliquée sur les trames avant de les additionner. Cette fenêtre doit assurer la conservation d'énergie.



(a) Signal synthétique obtenu par codec G.729
(b) Signal synthétique obtenu par codec G.729 avec perte.
(c) Signal synthétique obtenu par codec G.729 avec perte et PLC.

Figure 6: Résultats obtenus sur le chiffre '2'

La référence [8] contient une description et une analyse détaillée de la méthode de synthèse fonctionnant dans le domaine des fréquences. Le résultat produit par l'opération OLA remplace le quart de longueur d'onde du signal précédent l'effacement. Le résultat de cette opération OLA est placé dans la partie postérieure du tampon tonal et remplace le dernier quart de période du tampon historique. La Figure 7 montre l'amélioration de la qualité de parole (PESQ) avec l'utilisation de la technique PLC basé sur ITU-T G711 Appendix I pour des pertes de Taux=5% à Taux= 20%.

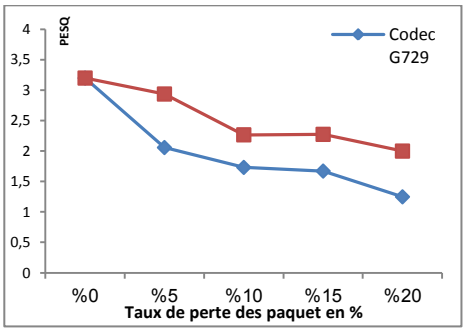


Figure 7. Evolution de la qualité de parole en fonction des pertes de paquets.

3. Reconnaissance automatique de la parole

La reconnaissance automatique de la parole est un processus qui convertit le signal acoustique de parole en un ensemble de mots ou de phrases. Les systèmes de RAP comportent les étapes suivantes :

3.1. Création de la base de données ARADIGIT_G729

La base de données parole originale utilisée dans ce travail est la base de données ARADIGIT. La base de données ARADIGIT8K (sous échantillonnée à 8 kHz) est passée ensuite à travers le Codec G.729 que nous avons simulé pour aboutir à la base de données transcodée G.729, à savoir la base nommée ARADIGIT_G729. La Figure 8 donne le processus de création de cette base.

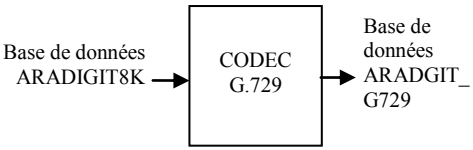


Figure 8: Création de la base de données ARADGIT_G729

3.2. Extraction des paramètres acoustiques

L'extraction des paramètres du signal consiste à associer au signal de parole une série de vecteurs de paramètres acoustiques. Cette étape consiste à le découper en trames pendant lesquelles il est supposé quasi-stationnaire, chaque trame a une durée de 25 ms, avec un recouvrement entre deux trames consécutives de 15 ms. Pour réduire les effets de bord produits par la segmentation, les trames sont alors multipliées par une fenêtre de pondération (la fenêtre de Hamming dans notre cas). À partir d'un banc de 24 filtres en échelle fréquentielle *Mel*, 12 paramètres MFCCs (*Mel-Frequency Cepstral Coefficients*) sont calculés pour chaque trame. A ces coefficients, des coefficients différentiels du premier et du second ordre sont ajoutés pour former un vecteur de dimension 36 (12 MFCC + 12 MFCC + 12MFCC).

3.3. L'étape d'apprentissage

Un modèle HMM à 3 états émetteurs est estimé pour chacun des mots (chiffres). La probabilité d'émission de chaque état

est modélisée par une distribution multi-gaussienne à matrice de covariance diagonale. HTK utilise l'algorithme de Viterbi pour initialiser les modèles prototypes ensuite l'algorithme de Baum-Welch pour les entraîner.

3.4. L'étape de reconnaissance

La reconnaissance est effectuée par l'algorithme de Viterbi qui calcule la vraisemblance entre la séquence d'observations acoustiques (le mot à reconnaître) et tous les modèles acoustiques ré-estimés dans l'étape de l'apprentissage. Le message reconnu est celui qui correspond au modèle acoustique qui engendre la vraisemblance maximale.

4. Résultats expérimentaux

Nous présentons dans cette partie les résultats de l'évaluation de l'influence des trames perdues sur la reconnaissance de la parole. Pour cela, nous utilisons la plateforme open source HTK basé sur les HMM. Pour minimiser l'influence de perte de paquet sur le taux de reconnaissance, nous utilisons le mécanisme de récupération des trames perdues PLC basé sur ITU-I G711 Appendix I.

Nous avons fait varier le taux des pertes de paquets de 0% jusqu'à 20%, pour les deux cas, c'est-à-dire avant et après l'application de technique du masquage des perte de paquets PLC basé ITU-T G711 Appendix I.

Pour un taux de perte allant de 0 jusqu'à 20%, l'influence des pertes sur le taux de reconnaissance est plus observable. Avec la technique PLC, l'influence des pertes sur le taux de reconnaissance est moins observable.

Globalement, le système de reconnaissance utilisé HTK basé sur les HMMs, prend comme référence lors de l'étape d'apprentissage un signal de parole sans perte. On remarque également que le signal reconstitué par la technique PLC est plus similaire au signal original que le signal perdu, même si la distorsion persiste et peut influer négativement sur la reconnaissance de parole. En comparant les résultats de la Figure 9, on peut noter que l'amélioration significative du taux de reconnaissance en utilisant la technique de masquage des pertes de paquets PLC basé sur ITU-T G711 Appendix I. Les résultats obtenus dans cette seconde phase ont montré que l'utilisation des techniques PLC améliore les performances de reconnaissance, en cas de perte de paquets. Les résultats obtenus avec notre méthode montrent un relèvement important du seuil de reconnaissance, donc l'efficacité de la méthode implémentée est significative.

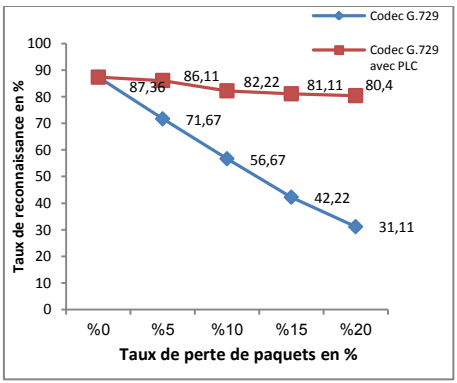


Figure 9 : Evolution du taux de reconnaissance en fonction des pertes de paquets

5. Conclusion

Dans ce travail, nous avons adapté la technique de dissimulation de la perte de paquets préconisée par la recommandation ITU-G711 Appendix 1 au codec G729 dédié à la VoIP. Notre principal objectif était l'amélioration de la RAP dans les réseaux VoIP. Nous avons pro-

posé l'introduction du PLC dans la reconnaissance dans les réseaux VoIP.
Les résultats expérimentaux montrent que notre méthode basée sur l'inclusion de la technique de dissimulation de perte peut être appliquée de manière efficace pour une application en reconnaissance vocale utilisant les réseaux VoIP. La solution proposée peut donc contribuer l'amélioration de la RAP en VoIP et rendre les systèmes de reconnaissance plus robustes quand aux pertes de paquets.

Il est évident qu'il faut ramener les résultats de notre travail aux spécifications des réseaux, notamment les protocoles utilisés.
Ainsi, une part non négligeable, et souvent prépondérante, dans les trames et paquets est constituée d'entêtes (Header). Il faut trouver une relation entre la perte de paquets transitant dans le réseau de communication et la portion effective de signal vocal. Ces travaux futurs sont envisagés dans le but de conforter les travaux sur la QoS, mais aussi les applications en reconnaissance vocale telles que la RAP et la RAL.

6. Références

[1] Young, S., Evermann, G., Kershaw, D., Moore, D., Odell, J., Ollason, D., Valtchev, V. and Woodland, P., "The HTK Book Version 3.3", Speech group, Engineering Department, Cambridge University. April 2005.

[2] ITU-T Recommandation G.729, "Codage de la parole à 8kbits/s par prédiction linéaire à excitation par séquences codées à structure algébrique conjuguée (CS-ACELP)", 1996.

[3] Yong, H. and Jiang, Z., "Implementation of ITU-T G729 Speech Codec in IP Telephony Gateway", Wuhan University Journal of Natural Sciences, vol. 5, pp.159-163, 2000.

[4] Milner, B. and Semnani, S. "Robust Speech Recognition Over Networks", IEEE International Conference Acoustics, Speech, and Signal Processing, Vol.3, pp. 1791 – 1794, 2000.

[5] Recommandation UIT-T G711, "Algorithme Simple de haut qualité pour le masquage des pertes en codage G.711", Septembre 1999.

[6] Wiley, J., "VoIP voice and fax signal processing", Published simultaneously in Canada, p.592, 2008.

[7] Sommen, P.C.W. and Jayasinghe, J.A.K.S., "On Frequency Domain Adaptive Filters using the Overlap-add Method", IEEE Philips Research Laboratories, pp.28-30, 1988.

[8] . Nakamura, K., "An Improvement of G.711 PLC Using Sinusoidal model", Proceedings of the IEEE The International Conference on Computer as a Toll, pp.1670-1673, 2005.