Modèles de Markov cachés appliqués au masquage de pertes de paquets en voix sur IP

Lionel KOENIG[1,2], Régine ANDRÉ-OBRECHT[1], Corinne MAILHES[1], Serge FABRE[2]

[1]IRIT - Université Paul Sabatier - ENSEEIHT, France

[2]Freescale semiconducteur, France

lionel.koenig@freescale.com, obrecht@irit.fr, corinne.mailhes@n7.fr, serge.fabre@freescale.com

Résumé − En téléphonie sur IP, le signal de parole est transmis sous forme de paquets sur le réseau. Afin de masquer les pertes de paquets et de reconstruire le signal de parole, nous présentons une méthode indépendante du codeur de parole s'appliquant directement sur le signal reconstruit après décodage. Cette méthode est basée sur une estimation à partir de modèles de Markov cachés (MMC). Un paramètre de voisement continu, appelé pourcentage de voisement, est introduit pour s'affranchir de la distinction voisée/non voisée et permet de n'avoir qu'un unique MMC. La méthode repose sur deux vecteurs de paramètres distincts : le premier associé à l'analyse du signal et à sa prédiction et le second adapté spécifiquement à la synthèse de la parole manquante. Les performances de ce nouvel indicateur de voisement ainsi que du système complet sont évaluées avec des résultats prometteurs.

Abstract − Packet loss due to misrouted or delayed packets in voice over IP leads to huge voice quality degradation. This paper presents a packet loss concealement algorithm which is independant from the vocoder. This method relies on hidden Markov model (HMM). A new voicing parameter is introduced to get over voiced/unvoiced sound separation and use a unique HMM. Since best parameter for prediction are not necessarily the best ones for synthesis, we introduce two separate vectors: the first one dedicated to the analysis of the signal and the second one featured for the synthesis of missing part. Performances of the proposed system are evaluated on parts of well-known speech corpora, leading to promising results.

1 Introduction

En téléphonie sur IP, la voix est transmise sous forme de paquets IP sur le réseau. Un paquet correspond au codage d'une trame de signal (environ $20ms$ de parole). En raison de l'architecture "Best-Effort" des réseaux IP, les paquets de voix peuvent être détruits lors d'une congestion, ou arriver suffisamment en retard pour être considérés par le système comme perdus. Le problème est donc de masquer les pertes. Plusieurs méthodes existent :

- l'*insertion de silence* n'apporte pas une qualité audio satisfaisante pour l'utilisateur.
- la *répétition* de paquets [4, 6]
- le masquage basé sur des *modèles de parole* plus coûteux en terme de calcul [5]

2 Masquage de perte de paquets basé sur des MMC

L'utilisation de modèles de Markov cachés (MMC) pour l'estimation de paquets perdus [11] est prometteuse. Rodbro et ses collègues s'appuient sur des MMC pour estimer un ou plusieurs paquets manquants : en supposant que toute suite de T paquets est associée à une suite de T vecteurs d'observations acoustiques $(\phi(t), t = 1, ...T)$, notée ϕ_1^T et supposée émise par un MMC acoustique, il est possible d'estimer (approche de type moindres carrés) une sous-suite manquante de L vecteurs, ϕ_t^{t+L-1}, connaissant les deux sous-suites passée ϕ_1^{t-1} et future ϕ_{t+L}^T . L'estimation de $\phi(t+k)$ pour tout $k = 0, ..., L-1$ est donnée par :

$$\hat{\phi}_{t+k} = \sum_{n=1}^{P} w_n \mu_n \tag{1}$$

où P est le nombre d'états du MMC, w_n est la probabilité contitionnelle d'être dans l'état n à l'instant $t+k$ connaissant les suites passée ϕ_1^{t-1} et future ϕ_{t+L}^T .

$$w_n = \Pr(s_{t+k} = n | \phi_1^{t-1}, \phi_{t+L}^{t+L+J-1}) \tag{2}$$

s est la variable aléatoire représentant les états du MMC, chaque état n représente un son acoustique élémentaire de type phone, et μ_n est la moyenne de la loi d'observation associée à cet état, loi supposée gaussienne. Dès lors que les paquets indexés $[t, t+1, ...t+L-1]$ sont manquants , l'équation (1) permet leur reconstruction.

Cependant, le système de Rodbro est directement lié au codeur de parole dans le sens où il utilise le codage de la trame t comme vecteur d'observation ϕ_t. Il conduit à

l'utilisation de deux MMC. En effet, deux des paramètres sont la période fondamentale et la fréquence de coupure de voisement ce qui conduit à un semi MMC composé d'un MMC pour traiter les sons voisés et un pour les sons non voisés. Notre étude reprend le cadre proposé par Rodbro en y apportant trois modifications majeures.

1. Le signal est analysé après décodage, afin d'être **indépendant du codeur**.
2. Pour n'utiliser qu'un **unique MMC**, nous avons étudié un nouvel indice de voisement continu que nous avons introduit dans le vecteur d'observation lui-même.
3. Une trame de signal est associée à deux types de vecteurs d'analyse, l'un est utilisé en phase d'estimation et l'autre en phase de synthèse, les meilleurs paramètres pour calculer les probabilités w_n (équation (2)) (étape de reconnaissance) ne sont pas forcément les meilleurs paramètres pour accéder à une très bonne synthèse de la trame manquante .

3 Le système de reconstruction de paquets manquants

L'architecture du système proposé est résumée dans le schéma 1.

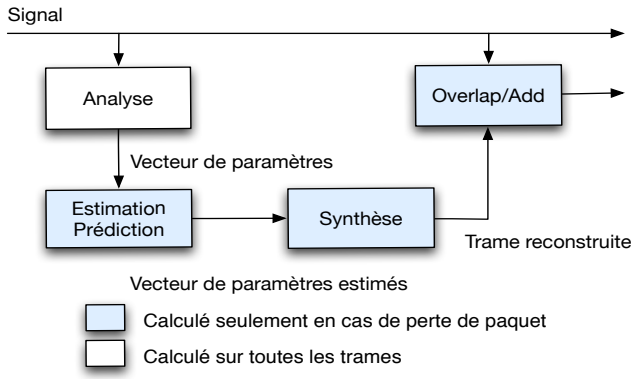


FIGURE 1 – Architecture du système de masquage de pertes.

Chaque trame t de signal est analysée pour donner le vecteur d'observation ϕ_t. Dès lors qu'une suite de paquets manque entre les instants t et $t+L-1$, une estimation des vecteurs manquants ϕ_t^{t+L-1} est produite à partir d'un MMC acoustique. À partir de ce traitement, une nouvelle suite de vecteurs ψ_t^{t+L-1} est estimée pour alimenter la phase de synthèse de la parole manquante.

3.1 Le modèle acoustique

Le modèle acoustique est un MMC de 256 états représentant des sons infraphonétiques ; chaque loi d'observation est une loi gaussienne. L'originalité de ce modèle provient du vecteur d'observation lui même. Compte-tenu de l'efficacité reconnue des coefficients cepstraux en reconnaissance de parole, nous avons utilisé les coefficients cepstraux issus de la prédiction linéaire (compromis entre efficacité du cepstre en décodage acoustique et proximité du codage). De plus, nous avons introduit une variable qui représente l'indice de voisement de la trame, qui se substitue à la détection voisé/non voisé et au suivi de la période fondamentale. Les coordonnées de l'observation ϕ_t sont finalement :

- L'énergie de la trame,
- Le pourcentage de voisement,
- 10 coefficients cepstraux issus d'un modèle auto-régressif d'ordre 10.

Pourcentage de voisement Le *pourcentage de voisement* $v_{\%}$ est défini comme le rapport entre la puissance voisée et la puissance totale [9] de la trame analysée. Une estimation de la puissance voisée du signal est obtenue en soustrayant la puissance du bruit à la puissance totale de la trame. La puissance du bruit est quant à elle obtenue en filtrant la densité spectrale de puissance (DSP) $S(\tilde{f})$ par un filtre médian :

$$v_{\%} = \frac{\int_0^{0.5} \Big(S(\tilde{f}) - \text{median}[S](\tilde{f}) \Big) \, \mathrm{d}\tilde{f}}{\int_0^{0.5} S(\tilde{f}) \mathrm{d}\tilde{f}} \qquad (3)$$

L'utilisation du filtre médian comme estimation de la DSP du bruit à partir de celle du signal bruité a été motivée par [2]. Dans cette thèse, plusieurs estimations ont été envisagées et l'utilisation du filtre médian semble un bon compromis en termes de complexité calculatoire et performances.

Mise en oeuvre et estimation des observations manquantes Le MMC, composé de 256 états, d'une loi gaussienne en dimension 12, est appris de manière classique sur le corpus OGI Multilingual Telephonic Speech [10]. Lors d'une perte de paquets, seules les pondérations w_n sont calculées à l'aide des deux suites ϕ_1^{t-1} et ϕ_{t+L}^{T}, en reprenant l'équation (2).

3.2 Synthèse vocale

Le procédé de synthèse vocale repose sur celui utilisé par Gunduzhan[5], qui nécessite une représentation spectrale. Afin d'obtenir une synthèse stable, nous avons choisi comme représentation spectrale les 10 premiers coefficients de type "Line Spectral Frequencies" (LSF), qu'il s'agit donc d'estimer lors d'une perte. Pour tout paquet manquant $t+k$, nous estimons ce vecteur ψ_{t+k} à partir des pondérations w_n (équation (2)) obtenues lors de la phase d'estimation :

TABLE 1 – Performances de segmentation

Corpus	Taux d'erreurs
OGI MLTS	17.0%
BREF80	10.7%

$$\hat{\psi}_{t+k} = \sum_{n=1}^{P} \Pr(s_{t+k} = n | \phi_1^{t-1}, \phi_{t+L}^{t+L+J-1}) \nu_n \qquad (4)$$

où les ν_n sont les vecteurs de type LSF correspondants aux μ_n, appris lors de la phase d'apprentissage.

4 Validation

4.1 Pourcentage de voisement

La validité du pourcentage de voisement a tout d'abord été vérifiée sur une tâche classique de segmentation automatique de la parole en zones voisées et en zones non voisées. Dans un deuxième temps, nous nous sommes intéressés à l'apport de ce nouveau descripteur de voisement en tant que paramètre d'un MMC. Cet apport a été validé à travers un décodeur acoustico-phonétique.

Ces deux évaluations ont été menées à la fois sur la partie étiquetée phonétiquement du corpus OGI Multilingual Telephonic Speech [10] représentant 11 heures de parole dans six langues et sur le corpus BREF80 [3] (11 heures de parole en français) sur lesquelles un alignement automatique a permis d'obtenir un étiquetage phonétique.

Segmentation voisé/non-voisé Une segmentation en zones voisées, silences et zones non-voisées est obtenue en seuillant à la fois le pourcentage de voisement et la puissance de la trame. Le seuil de segmentation est positionné de façon à minimiser le taux d'erreurs sur l'ensemble du corpus. Cette segmentation est ensuite comparée à celle obtenue à l'aide d'un algorithme de détection automatique de fréquence fondamentale [1]. Les resultats de cette comparaison en termes de taux d'erreurs sont présentés dans le tableau 1. Ils montrent que le pourcentage de voisement décrit correctement le trait de voisement, les taux d'erreurs étant tout-à-fait acceptables.

Toutefois, cette première validation classique d'un taux de voisement ne suffit pas pour assurer l'intérêt de l'introduction de ce paramètre dans un MMC. C'est pourquoi une deuxième validation a été envisagée permettant de conclure sur l'apport de ce paramètre dans un MMC. Cette deuxième validation est présentée ci-après.

Décodeur acoustico-phonétique L'ajout d'un paramètre non-homogène à un vecteur d'observation de modèle de Markov peut conduire à une importante baisse des performances. Nous avons donc comparé les performances d'un système classique à celles d'un système similaire incluant le pourcentage de voisement dans son vecteur de paramètres. La paramétrisation utilisée comporte l'énergie, les douze premiers coefficients cepstraux issus de la prédiction linéaire avec soustraction de moyenne cepstrale ainsi que leurs dérivées. La précision et le taux de reconnaissance phonétique (TRP) des deux modèles sont présentés dans le tableau 2.

TABLE 2 – Taux de reconnaissance phonétique

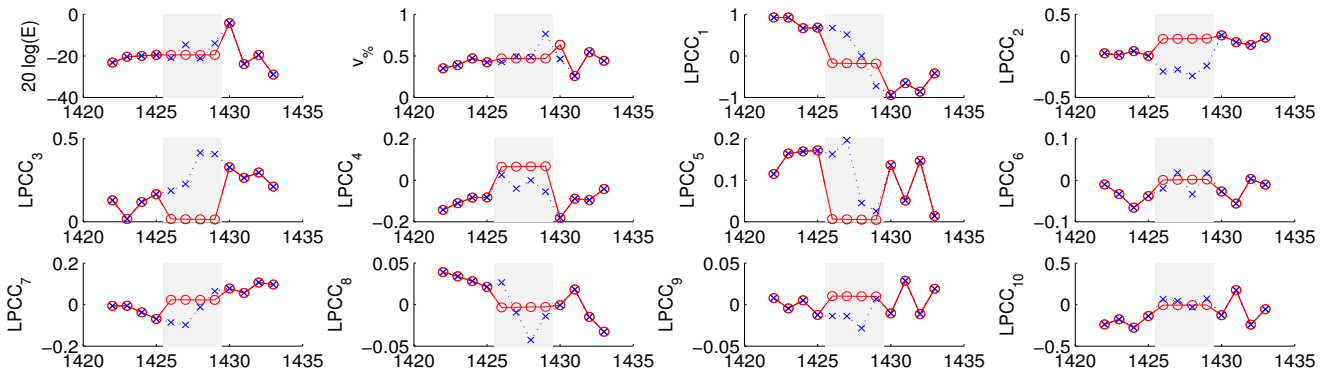
Modèle	Précision	TRP
Décodeur classique LPCC_E_D_Z	59.92(±0.5)%	67.92(±0.5)%
Décodeur proposé LPCC_E_D_Z + $V_{\%}$	60.39(±0.5)%	68.42(±0.5)%

On remarque que l'ajout du paramètre de voisement dans le MMC non seulement ne dégrade pas les performances du système de décodage mais il apporte même une légère augmentation dans les résultats obtenus. Ceci nous suffit pour conclure que ce paramètre a sa place dans un MMC et qu'il pourra être utilisé en tant qu'un paramètre caractérisant la part de voisement contenu dans une trame d'un signal de parole.

4.2 Validation du système global

Ayant validé l'intérêt de rajouter ce paramètre de voisement dans le vecteur paramètre du MMC, nous pouvons tester le système proposé de reconstruction de paquets manquants. L'avantage d'introduire un paramètre continu de voisement est d'éviter l'utilisation de deux MMC comme le fait Rodbro (l'un pour les sons voisés, l'autre pour les non voisés). La validation du système global s'est faite en utilisant le corpus OGI [10].

Nous avons simulé des pertes de paquets ($L < 10$) et nous avons mesuré la qualité du signal de sortie avec le Perceptual Evaluation of Speech Quality (PESQ) [7] entre le signal d'origine et le signal après pertes et reconstitution de paquets. Les résultats présentés dans le tableau 3 permettent de comparer les résultats obtenus par le système proposé (dénommé HMM) à ceux issus d'autres systèmes comme l'insertion de silence ou la répétition de paquets dans le G711. Ces résultats semblent indiquer que le système proposé fournit des performances supérieures à l'insertion de silence mais inférieures à la répétition de paquets. Néanmoins, la qualité du signal de sortie (mesurée par le PESQ) est en grande partie liée à celle du synthétiseur de parole. Or, dans nos simulations, nous ne pouvons assurer que le synthétiseur utilisé soit de très grande qualité, ce qui suffit à dégrader les résultats mesurés par le PESQ sur notre système. C'est pourquoi, nous donnons sur la figure 2 un exemple des vecteurs estimés dans le cas d'une zone transitoire. Les distances observées à la fois sur l'énergie, le paramètre de voisement et la distance



en ligne continue -x- le paramètre original, en pointillés -o- le paramètre reconstruit.

FIGURE 2 – Vecteur estimé dans le cas d'une perte de paquets en zone transitoire ($L = 4$ et $J = 3$)

TABLE 3 – Scores PESQ

Corpus	**Taux pertes**	**Silence insertion**	**G711** *Appendix 1*	**HMM**
OGI	1 %	3.84	4.07	3.87
	5 %	2.86	3.38	2.91
MLTS	10 %	2.24	2.92	2.30

cepstrale sont de faibles valeurs comparées aux distances entre deux états du HMM, comme le montre l'étude [8], ce qui confirme la qualité du masquage de paquets proposé.

5 Conclusions

Nous avons présenté un système de masquage de pertes de paquets basé sur des MMC. La contribution des deux vecteurs de paramètres le premier pour l'analyse et le second pour la synthèse permet d'avoir une paramétrisation adaptée à chaque tâche. Nous projetons d'explorer plus en avant le choix des deux paramétrisations ainsi que la difficile tâche de mesure de performances du système de masquage de pertes. En effet, comme nous souhaitons être indépendant du codeur, le synthétiseur n'est pas fixé a priori. Or la mesure classique de performances de la qualité d'un signal parole est le PESQ et celui-ci dépend fortement de la qualité du synthétiseur utilisé. Dans notre étude, nous souhaitons mesurer la qualité du système de masquage de paquets proposé, et non pas la qualité du synthétiseur. C'est pourquoi le PESQ ne semble pas la mesure appropriée dans notre cas et d'autres mesures doivent être envisagées. Par exemple, des distances calculées sur les paramètres du HMM pourraient être mises en oeuvre mais avec le problème de la non-homogénéité des composantes de ce vecteur.

Références

[1] A. de CHEVEIGNÉ et H. KAWAHARA : YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111:1917, April 2002.

[2] M. DURNERIN : *Une stratégie pour l'interprétation en analyse spectrale, détection et caractérisation des composantes d'un spectre*. Thèse de doctorat, INPG, 1999.

[3] J.-L. GAUVAIN, L. LAMEL et M. ESKÉNAZI : Design considerations and text selection for BREF, a large french read-speech corpus. *In International Conference on Speech and Language Processing*, vol. 2, p. 1097–2000, Kobe, Japan, nov. 1990.

[4] D. GOODMAN, O. JAFFE, G. LOCKHART et W. WONG : Waveform substitution techniques for recovering missing speech segments in packet voice communications. *In Proc. IEEE ICASSP*, vol. 11, p. 105–108, 1986.

[5] E. GUNDUZHAN et K. MOMTAHAN : Linear prediction based packet loss concealment algorithm for PCM coded speech. *IEEE Trans. on Speech and Audio Processing*, 9(8):778–785, 2001.

[6] ITU RECOMMANDATION G.711 : Pulse code modulation (PCM) of voice frequencies. ITU Recommendation G.711, ITU Recom., Nov 1988.

[7] ITU-T STUDY GROUP 12 : Perceptual evaluation of speech quality (PESQ). ITU Recommendation P.862, ITU Recom., Feb 2001.

[8] L. KOENIG : Validation du masquage de paquets basé sur un hmm. Rapport interne, Institut de recherche en informatique de Toulouse, 2009.

[9] L. KOENIG, C. MAILHES, R. ANDRÉ-OBRECHT et S. FABRE : A continuous voicing parameter in the frequency domain. *In 13th International Conference on Speech and Computer*, 2009.

[10] Y. K. MUTHUSAMY, R. A. COLE et B. T. OSHIKA : The OGI multilanguage telephone speech corpus. *In Proc. of Int. Conf. on Speech and Language Processing*, p. 895–898, Oct 1992.

[11] C. RODBRO, M. MURTHI, S. ANDERSEN et S. JENSEN : Hidden Markov model-based packet loss concealment for voice over IP. *IEEE Trans. on Audio, Speech and Language Processing*, 14(5):1609–1623, 2006.