Évaluation des systèmes d'analyse-modification-synthèse de parole

Gérard Bailly

Institut de la Communication Parlée - UMR CNRS n^o5009 - INPG & Université Stendhal
46, avenue Félix Viallet, 38031 Grenoble Cedex 1, France
Tél. : ++33 04 76 57 47 11 - Fax : ++33 04 76 57 47 10
e-mail : bailly@icp.inpg.fr - http ://www.icp.inpg.fr/

Abstract

The evaluation of analysis-modification-synthesis systems provided in the literature is often informal and results cannot be compared. We propose here a methodology for comparing the performance of such systems using common ressources. The proposed benchmark consists in performing various prosodic transplantations from a given natural utterance to various prosodic "variants" of the same sentence. These prosodic "variants" should be recorded by the same speaker in order to obtain both credible prosodic transplantations and reference target signals. These principles and first results obtained within the framework of the Cost258 project are presented.

Introduction

La plupart des systèmes de synthèse de parole utilise la concaténation d'unités stockées pour générer un continuum sonore dont certaines caractéristiques seront alors modifiées en fonction de consignes prosodiques (voir figure 1). Le travail est donc partagé entre l'algorithme de sélection des unités et le système d'analyse-modification-synthèse de signal (appelé "codeur" dans la suite). Bien qu'un algorithme de sélection opérant sur de larges dictionnaires d'unités peut réduire la distorsion à effectuer par le codeur en incorporant dans son processus de sélection la minimisation de la distance des caractéristiques prosodiques des signaux sélectionnés à la cible désirée [6, 15], il est cependant difficile d'obtenir une qualité de voix homogène sur de larges corpus et il semble difficile d'éviter le recours à un codeur ne serait-ce que pour effectuer un mise à l'échelle ou un lissage de certaines caractéristiques spectrales.

La phase de modification des caractéristiques prosodiques est cruciale : la modification d'une caractéristique aussi standard que la mélodie est accompagnée dans les signaux naturels d'une covariation de l'ensemble de la représentation paramétrique du signal. Ainsi les variations de la forme de l'onde glottique ou des formants en fonction de la fréquence fondamentale ou de l'effort vocal ont été bien étudiés [11] et utilisés par les systèmes de synthèse par règles. Dans le cas de codeurs sans accès explicite à ces paramètres fins, les covariations doivent être implicites. Parmi ces propriétés implicites, on trouve par exemple l'invariance de forme, propriété commune à de nombreux codeurs (TDPSOLA [7], codeurs sinusoïdaux [1, 16] ...) qui préserve la forme globale de l'onde malgré les changements de fréquence fondamentale. La préservation de ces propriétés implicites n'est cependant ni nécessaire ni suffisante pour garantir la cohérence des modifications et il semble difficile pour l'instant de s'affranchir d'une évaluation comparative des codeurs.

Cet article propose une méthodologie d'évaluation des codeurs de synthèse basée sur des tâches de transplantation prosodique de complexité croissante.

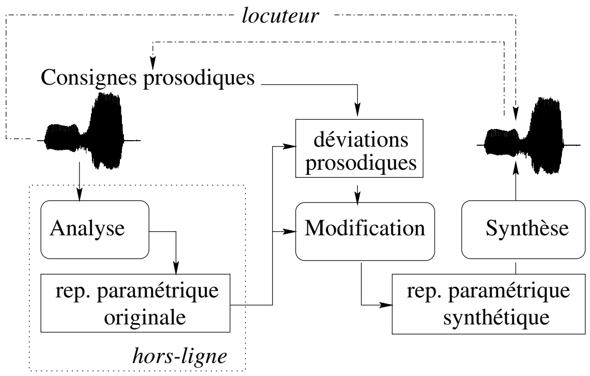

Fig. 1: Schéma de principe d'un codeur. Grâce à une méthode d'analyse spécifique, on délivre une représentation de signaux originaux qui sera modifiée pour obéir aux consignes prosodiques calculées par le processus de synthèse. Le principe d'évaluation proposé est figuré en traits discontinus : un même locuteur prononce les signaux sources et cibles. Le codeur est alors soumis à une tache de transplantation prosodique.

1. Évaluation des codeurs : un problème mal-posé

Du fait de l'émergence des méthodes statistiques en synthèse de parole, de la (relative) facilité de mise en oeuvre de systèmes de synthèse par concaténation, un système de synthèse n'est plus nécessairement le fruit d'un travail acharné de quelques experts phonéticiens. Grâce à des outils génériques [10, 20] et des ressources de qualité disponibles, il est facile de construire des systèmes opérationnels. On comprend dès lors la nécessité d'élaborer des méthodologies d'évaluation modulaire ("glass box") afin d'identifier les propriétés intrinsèques des briques du lego communautaire. Et la littérature ne nous fournit que peu d'expériences de ce type : la plupart des évaluations impliquent des systèmes complets (d'ailleurs souvent identifiés par le codeur utilisé [17]) ou comparent les codeurs au sein d'une même architecture [9, 18, 19]. Outre le fait que ces comparaisons n'incluent pas une échelle de référence (par exemple, de la parole naturelle ou au moins de la prosodie naturelle), il est impossible d'établir une grille de lecture des résultats tant les propriétés des codeurs sont différentes en regard des tâches qui leur sont demandées : ainsi TDPSOLA est très sensible aux problèmes de concaténation (phases, structure spectrales aux points de concaténation) mais permet de préserver avec grande précision des zones du signal difficiles à analyser avec des codeurs paramétriques [5]. D'autre part, les propositions de codeurs ne sont accompagnées le plus souvent que d'évaluation informelle impliquant les codeurs dans des tâches peu réalistes (manipulation de durées ou de fréquence fondamentale uniforme de signaux naturels) sur des stimuli ad hoc [1]. L'auditeur est ainsi biaisé vers un jugement esthétique sans référence aux performances attendues.

Les évaluations subjectives sont lourdes - et donc précieuses - et présentent une photographie d'un système à un instant donné de son développement. Il y a donc nécessité de capitaliser les résultats obtenus de manière à permettre aux chercheurs d'utiliser l'évaluation comme outil de diagnostic et d'enrichir l'éventail des tâches de modification et des méthodologies d'évaluation afin de mettre en valeur les propriétés attendues des codeurs : un système de synthèse par concaténation de larges unités ayant peu recours à la modification insistera sur la transparence du processus d'analyse-synthèse alors que des systèmes de synthèse multi-style faisant largement appel à des modèles de variabilité mettra l'accent sur la versatilité du codeur.

2. Identification de tâches

Alors que la recherche en reconnaissance de parole, identification de la langue et du locuteur est largement structurée et dynamisée par des campagnes d'évaluation internationales, l'évaluation des systèmes de synthèse reste un sujet controversé. Peu de campagnes d'évaluation permettent d'effectuer un réel diagnostic des principales lacunes des modules utilisés dans les synthétiseurs actuels ou de leur mode de couplage. Notons cependant l'efficacité d'une telle campagne lorsque les bonnes grilles de lecture des résultats sont identifiées [22].

2.1. Tâches élémentaires

La grille de lecture des performances des codeurs proposée s'appuie sur l'identification de tâches élémentaires auxquelles les codeurs sont soumis et sur lesquelles une différenciation nette est attendue. Une liste non exhaustive comprendra d'abord la manipulation des paramètres prosodiques classiques : la fréquence fondamentale, la durée et l'intensité des sons, puis la manipulation du timbre pour diverses applications telles que lissage aux points de concaténation, génération multi-style ou transformation de voix. Il convient de plus d'évaluer des manipulations sur divers types de sons : ainsi, la manipulation du fondamental de fricatives voisées pose le problème de la synchronisation entre bruit de friction et oscillation glottique [14] ; de même la manipulation des durées de sons non-voisés pose le problème de la définition et du traitement des trames à beaucoup de codeurs synchrones avec le fondamental.

2.2. Vers des tâches complexes

Comme on l'a vu dans l'introduction, ces manipulations ne sont pas totalement contrôlées de manière indépendante et doivent être effectuées de manière synergétique pour structurer le discours. Une évaluation doit donc comprendre une complexification progressive des tâches de manière à tester non plus le module d'analyse-synthèse mais aussi le module de modification des représentations qui prend en charge la négociation des consignes.

2.3. Ressources

Signaux La comparaison des codeurs à une référence absolue naturelle suppose de disposer de ressources source et cible illustrant ces manipulations élémentaires et complexes. Le locuteur doit ainsi être instruit des tâches soit *de manière implicite* par usage de stimuli synthétiques guidant ses réalisations soit directement par imitation soit indirectement en donnant des buts à atteindre (Barbosa [4] a par exemple suggéré des variations de débit par des questions synthétiques à la place du métronome souvent utilisé) ; soit *de manière explicite* par description textuelle d'une situation propre à suggérer une intonation voulue.

Descripteurs Il convient de disposer de descripteurs prosodiques de référence afin de pouvoir comparer les codeurs dans les même tâches de transplantation prosodique. Ces descripteurs doivent donc au minimum comprendre : un marquage de périodes, un marquage des frontières phonémiques. Ils peuvent et doivent ensuite être enrichis par des descripteurs "intelligibles" i.e. sur lesquels une connaissance explicite puisse être acquise (pente spectrale, coefficients de description de l'onde glottique, jitter, shimmer ...).

3. L'exemple du serveur Cost258

Dans le cadre du Cost258 [2], une première version d'un serveur de ressources [3] pour l'évaluation des codeurs a été mis en place.

3.1. Les ressources

Plusieurs locuteurs ont rempli les tâches suivantes : (1) VO (*hauteurs vocaliques*) : les 10 voyelles orales du Français ont été prononcées à plusieurs hauteurs de manière à couvrir une plage d'environ 3/4 d'octave ; (2) FD (*durées*) : deux versions courtes et longues des 6 fricatives du Français ont été produites isolément. (3) AT (*attitudes*) : 6 phrases ont été pronon-

[1] rendons grâce à Speech Communication qui permet de proposer en ligne les stimuli (`www.elsevier.nl :80/inca/publications/store/5/0/5/5/9/7` utilisés par les auteurs (voir par exemple [21])

[2] `www.unil.ch/imm/docs/LAIP/COST_258/cost258.htm`

[3] `www.icp.inpg.fr/cost258/evaluation/server/... cost258_coders.html`

cées avec 6 attitudes prosodiques différentes. (4) EM (*émotion*) : 1 phrase a été prononcée avec 4 émotions différentes.

Chaque locuteur a aussi prononcé chaque énoncé avec une intonation et un rythme les plus plats possible, de manière à fournir un stimuli source le plus proche possible de la base fournie par un système par concaténation idéal (voir par exemple la normalisation opérée par le système MBROLA). Tous les signaux ont été échantillonnés à 16kHz, segmentés et les marqueurs de périodes positionnés semi-automatiquement. Le centre de réalisation des phonèmes a été marqué et l'énergie à court-terme de ce centre a été ajouté au jeu de descripteurs prosodiques.

3.2. Les codeurs de référence

Plusieurs codeurs ont accompli les diverses tâches de transplantation : nous avons implémenté une version éprouvée de TDPSOLA [3] ainsi qu'une alternative appelée LPPSOLA où TDPSOLA opère sur le résidu de prédiction linéaire. 4 autres codeurs fournis par d'autres laboratoires partenaires ont traité ces stimuli. Trois autres codeurs ont été simulés en introduisant divers niveaux de bruit additionnel (+30dB, +20dB, +10dB).

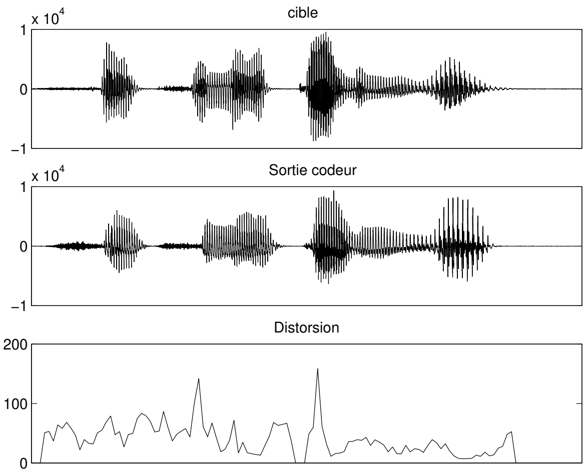

Fig. 2: Évolution de l'erreur de modélisation par la mesure de distorsion WSS.

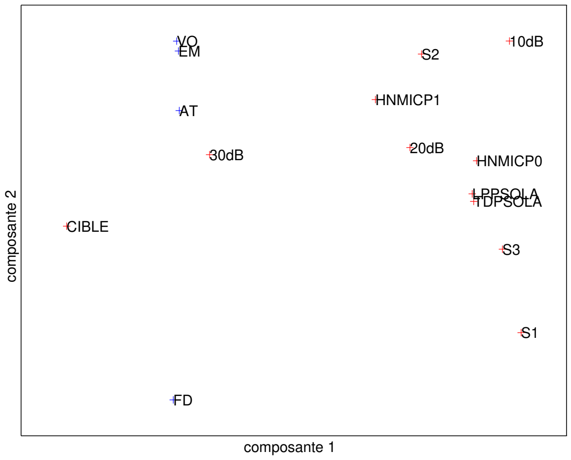

Fig. 3: Projection de 7 codeurs sur le premier plan factoriel obtenu par ACP des distorsions moyennes qu'ils obtiennent sur les diverses tâches de transplantation.

4. Procédures d'évaluation

Un tel serveur vise à fournir des ressources de référence aux développeurs de codeurs. Il vise aussi à fournir une base méthodologique d'évaluation des systèmes.

4.1. Évaluation objective

De nombreux travaux s'attachent à reproduire par des mesures de distorsion les scores moyens d'opinion (MOS) obtenus lors d'évaluations subjectives telles que l'estimation de la gêne perceptive occasionnée par des discontinuités spectrales aux points de concaténation entre deux segments de parole [8, 15] ou la prédiction plus globale des performances d'algorithmes de réhaussement de parole [12].

Mesures de distorsion Les nombreuses mesures de distorsion utilisées dans ces études estiment la distance ou la corrélation existant entre les représentations temps-fréquence des signaux comparés. Elles sont sensibles à diverses caractéristiques et il est impossible de sélectionner une seule et unique mesure capable de reproduire les jugements divers demandés aux sujets. De plus, ces mesures délivrent une mesure dynamique (voir Figure 2) qu'il est difficile de relier à une MOS globale (voir cependant [13]) : les mesures sont donc souvent moyennées.

Carte d'évaluation globale Au lieu de choisir une seule mesure, nous avons laissé à un algorithme de post-traitement le soin de "choisir" parmi un jeu de distorsions une combinaison permettant de différencier au mieux les résultats des codeurs. Suivant les propositions de Hansen et Pellom [12], nous avons calculé plusieurs mesures de distorsions centisecondes, entre cibles et transplantations prosodiques, utilisant : la pente spectrale pondérée (WSS), le quotient de similitude logarithmique (LLR) où les rapports d'aires logarithmiques (LAR) [4]. Chaque codeur est ainsi caractérisé par a jeu de 90 distorsions moyennes (3 mesures x 15 tâches x 2 caractéristiques (moyenne, écart-type)). 6 systèmes ont été testés : TDPSOLA, LPPSOLA, la version originale HNMICP0 d'un codeur harmonique+bruit [2], une version améliorée suite au test (HNMICP1) et trois autres codeurs (S1,S2,S3). La cible bruitée à divers niveaux de bruit a été ajoutée (rapport signal sur bruit (RSB) de 10dB, 20dB et 30dB).

De manière à produire une représentation qui reflète la distance globale de chaque codeur à la cible et qui maximise leurs différences, ce jeu de 9x90 distorsions a été projeté sur le premier plan factoriel (voir Figure 3). Les trois premières composantes de l'analyse factorielle normalisée expliquent respectivement 87.7, 7.3 et 2.9 % de la variance des mesures de distorsion. Nous avons projeté de même les moyennes obtenues par tous les codeurs sur chaque corpus, les points sont

[4]Cette liste n'est pas limitative et ne change en rien le principe.

figurés par (AT,EM,VO,FD).

Commentaires Globalement les codeurs se situent à une dégradation équivalente de 10 à 20dB de RSB. On ne trouve pas de différence significative entre les systèmes TDPSOLA et LPPSOLA. L'amélioration des performances du système HNMICP correspond bien à un rapprochement de la projection du système vers la cible. L'évolution de la projection des signaux bruités et le placement des diverses projections des corpus sur le plan factoriel permettent d'interpréter le premier axe factoriel comme une mesure du RSB, le deuxième comme une distorsion au voisement - ce qui explique pourquoi un RSB de 10dB a une ordonnée plus forte que les autres RSB : les mesures de distorsion étant sensibles aux formants, lorsque ceux-ci sont noyés dans le bruit, la distorsion augmente très rapidement.

Conclusions

Il est facile de mettre à disposition des ressources sous format électronique permettant une évaluation cumulative et modulaires des systèmes de synthèse en vue d'un diagnostic aussi bien que d'une évaluation comparative. Nous espérons compléter ce serveur suivant trois directions : (1) augmenter l'éventail des descripteurs prosodiques des stimuli, (2) valoriser les propriétés des codeurs grâce à l'introduction de nouvelles tâches, notamment de concaténation de stimuli ou de manipulation de timbre, (3) mettre à disposition des méthodologies d'évaluation objective et cumuler les résultats d'évaluation subjective opérés sur ces stimuli. Nous espérons que la communauté francophone nous aidera dans cette démarche.

Remerciements

Ce travail a été supporté par le Cost258 et l'ARC-B3 de l'AUPELF-UREF. Grand merci à Eric Keller, Alex Monaghan, Eduardo Rodríguez Banga et Erhart Rank.

Références

[1] Almeida, L.B. and Silva, F. Variable-frequency synthesis : an improved harmonic coding scheme. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 27–5, 1984.

[2] Bailly, G. Accurate estimation of sinusoidal parameters in an harmonic+noise model for speech synthesis. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 3, pages 1051–1054, 1999.

[3] Bailly, G., Barbe, T., and Wang, H. Automatic labelling of large prosodic databases : tools, methodology and links with a text-to-speech system. In Bailly, G. and Benoît, C., editors, *Talking Machines : Theories, Models and Designs*, pages 323–333. Elsevier B.V., 1992.

[4] Barbosa, P. *Caractérisation et génération automatique de la structuration rythmique du français*. Thèse de troisième cycle, Institut National Polytechnique de Grenoble, Grenoble, France, 1994. Thèse de Doctorat Spécialité Sciences Cognitives sous la direction de Gérard Bailly.

[5] Böeffard, O. and Violaro, F. Improving the robustness of text-to-speech synthesizers for large prosodic variations. In *ETRW on Speech Synthesis*, pages 111–114, New Paltz - New York, 1994.

[6] Campbell, W.N. Synthesizing spontaneous speech. In Sagisaka, Y., Campbell, N., and Higuchi, N., editors, *Computing prosody : Computational models for processing spontaneous speech*, pages 165–186. Springer Verlag, 1997.

[7] Charpentier, F. and Moulines, E. Pitch-synchronous waveform processing techniques for text-to-speech using diphones. *Speech Communication*, 9(5-6) :453–467, 1990.

[8] Ding, W., Fujisawa, K., and Campbell, N. Improving speech synthesis of CHATR using a perceptual discontinuity function and constraints of prosodic modification. In *ETRW Workshop on Speech Synthesis*, pages 191–194, Jenolan Caves - Australia, 1998.

[9] Dutoit, T. High quality text-to-speech synthesis : A comparison of four candidate algorithms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 565–568, Adelaide - Australia, 1994.

[10] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and van der Vrecken, O. The MBROLA project : towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proceedings of the International Conference on Speech and Language Processing*, volume 3, pages 1393–1396, Philadelphia - USA, 1996.

[11] Gobl, C. and Chasaide, N. Acoustic characteristics of voice quality. *Speech Communication*, 11(4-5) :481–490, 1992.

[12] Hansen, J.H. and Pellom, B.L. An effective quality evaluation protocol for speech enhancement algorithms. In *Proceedings of the International Conference on Speech and Language Processing*, volume 6, pages 2819–2822, 1998.

[13] Hansen, M. and Kollmeier, B. Continuous assessment of time-varying speech quality. *Journal of the Acoustical Society of America*, 105(5) :2888–2899, 1999.

[14] Hermes, D.J. Synthesis of breathy vowels : Some research methods. *Speech Communication*, 10 :497–502, 1991.

[15] Klabbers, E. and Veldhuis, R. On the reduction of concatenation artefacts in diphone synthesis. In *Proceedings of the International Conference on Speech and Language Processing*, volume 5, pages 1983–1986, 1998.

[16] Quatieri, T.F. and McAulay, R.J. Phase coherence in speech reconstruction for enhancement and coding applications. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 207–210, 1989.

[17] Sonntag, G.P., Portele, T., Haas, F., and Köhler, J. Comparative evaluation of six German TTS systems. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 1, pages 251–254, Budapest - Hungary, 1999.

[18] Stylianou, Y. Concatenative speech synthesis using a harmonic plus noise model. In *3rd ESCA/COCOSDA International Workshop on Speech Synthesis*, pages 261–266, Jenolan Caves, Australia, november 1998.

[19] Syrdal, A.K., Möhler, G., Dusterhoff, K., Conkie, A., and Black, A.W. Three methods of intonation modeling. In *3rd ESCA/COCOSDA International Workshop on Speech Synthesis*, page accepted, Jenolan Caves, Australia, november 1998.

[20] Taylor, P., Black, A.W., and Caley, R. The architecture of the FESTIVAL speech synthesis system. In *Third International Workshop on Speech Synthesis*, pages 147–151, Jenolan Caves, Australia, 1998.

[21] Veldhuis, R. and Hé, H. Time-scale and pitch modifications of speech signals and resynthesis from the discrete short-time Fourier transform. *Speech Communication*, 18 :257–279, 1996.

[22] Yvon, F., de Mareuil, P.B., d'Alessandro, C., Aubergé, V., Bagein, M., Bailly, G., Béchet, F., Foukia, S., Glodman, J.F., Keller, E., O'Shaughnessy, D., Pagel, V., Sannier, F., Véronis, J., and Zellner, B. Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in French. *Computer Speech and Language*, 12 :393–410, 1998.