Se concentrer sur les différences : une méthode d'évaluation subjective efficace pour la comparaison de systèmes de synthèse

Jonathan Chevelu[1] Damien Lolive[1] Sébastien Le Maguer[2] David Guennec[1]
(1) IRISA, Université de Rennes 1, Lannion, France
(2) Saarland University, Saarbrücken, Germany
jonathan.chevelu@irisa.fr, damien.lolive@irisa.fr,
david.guennec@irisa.fr, slemaguer@coli.uni-saarland.de

RÉSUMÉ

En proposant une nouvelle approche de synthèse de la parole, les études comportent généralement une évaluation subjective d'échantillons acoustiques produits par un système de référence et un nouveau système. Ces échantillons sont produits à partir d'un petit ensemble de phrases choisies aléatoirement dans un unique domaine. Ainsi, statistiquement, des échantillons pratiquement identiques sont présentés et réduisent les écarts de mesure entre les systèmes, au risque de les considérer comme non significatifs. Pour éviter cette problématique méthodologique, nous comparons deux systèmes sur des milliers d'échantillons de différents domaines. L'évaluation est réalisée uniquement sur les paires d'échantillons les plus pertinentes, c'est-à-dire les plus différentes acoustiquement. Cette méthode est appliquée sur un système de synthèse de type *HTS* et un second par sélection d'unités. La comparaison avec l'approche classique montre que cette méthode révèle des écarts qui jusqu'alors n'étaient pas significatifs.

ABSTRACT

Focus on differences : a subjective evaluation method to efficiently compare TTS systems *.

When trying to assess the effectiveness of a new speech synthesis method, researchers usually conduct subjective evaluations by randomly choosing a small set of samples, from the same domain, taken from a baseline system and the proposed one. When selecting them randomly, statistically, samples with almost no differences are evaluated and the global measure is smoothed which may lead to judge the improvement not significant. To solve this methodological flaw, we propose to compare speech synthesis systems on thousands of generated samples from various domains and to focus subjective evaluations on the most relevant ones by computing a normalized alignment cost between sample pairs. This process has been successfully applied both in the HTS statistical framework and in the unit selection approach. A comparison between tests involving most different samples and randomly chosen samples shows clearly that the proposed approach reveals significant differences between the systems.

MOTS-CLÉS : synthèse de la parole, évaluation subjective, différence acoustique.

KEYWORDS: speech synthesis, subjective evaluation, acoustic difference.

∗. Cet article reprend un travail présenté par les mêmes auteurs à la conférence Interspeech 2015.

1 Introduction

Dans le domaine de la synthèse de la parole (TTS), l'évaluation subjective est cruciale puisque l'objectif principal est de produire un message destiné à des auditeurs humains. Des évaluations objectives et subjectives peuvent être utilisées. D'une part, les évaluations objectives ont l'avantage d'être peu coûteuses à réaliser, mais peu importe leur précision, elles ne peuvent pas encore remplacer les tests subjectifs. D'autre part, pour être intéressantes, les évaluations subjectives ont besoin d'un grand nombre d'auditeurs et d'un grand nombre d'échantillons choisis en fonction du domaine du contexte d'utilisation de la synthèse.

Plusieurs types d'évaluations perceptives sont généralement utilisées. Parmi toutes ces méthodes, on peut distinguer des tests de préférence comme AB et ABX, des tests de pointage comme MOS (*Mean Opinion Score*), DMOS (*Degradation MOS*) ou plus récemment MUSHRA (*MUltiple Stimuli with Hidden Reference and Anchor*). Toutes ces méthodes ont le même objectif, à savoir le classement des systèmes selon certains critères subjectifs.

Dans la littérature, la plupart des propositions scientifiques sont évaluées à l'aide de tests perceptifs mais le nombre d'échantillons étudiés reste très limité. Par exemple, le défi Blizzard est composé de campagnes d'évaluation à grande échelle (King & Karaiskos, 2012; Prahallad *et al.*, 2014) mais ne comporte que quelques centaines de signaux. Ceci se retrouve dans d'autre travaux, parmi lesquels nous pouvons citer (Sainz *et al.*, 2014) avec 350 phrases, (Garcia *et al.*, 2006) avec 7 phrases pour 5 systèmes ou encore (Hinterleitner *et al.*, 2011) avec deux groupes de 18 stimuli. Ce faible nombre de stimuli est généralement motivé par l'aspect particulièrement chronophage des campagnes d'évaluation perceptives. Quelques travaux récents ont mis en doute la méthodologie d'évaluation, comme (Latorre *et al.*, 2014) qui étudie l'impact des références mentales des auditeurs sur les résultats des tests perceptifs, ou (Hinterleitner *et al.*, 2011; Viswanathan & Viswanathan, 2005) qui ont proposé des modifications des protocoles existants. Des alternatives aux méthodes classiques ont également été utilisées, sur le principe d'évaluations en ligne à grande échelle (*crowdsourcing*) comme décrit dans (Buchholz *et al.*, 2013).

Plus important que le petit nombre d'échantillons choisis, le fait qu'ils soient choisis au hasard et non pas pour leur importance pour les méthodes d'évaluation peut biaiser les résultats des évaluations. Dans cet article, contrairement à ce qui se fait habituellement, nous proposons de synthétiser un grand nombre d'échantillons (plusieurs milliers), à partir de textes de divers domaines. Compte tenu du nombre élevé d'échantillons, nous introduisons un coût d'alignement entre les paires d'échantillons provenant de deux systèmes différents afin de les classer par similarité acoustique. Une fois cela fait, nous pouvons construire une évaluation perceptive en utilisant uniquement les signaux les plus différents. De cette façon, nous ne faisons aucune hypothèse concernant la qualité d'un système parmi les autres, nous concentrons simplement l'évaluation sur ce qui peut faire la différence entre les systèmes. Une telle stratégie permet de réduire la taille d'une évaluation perceptive pour recentrer l'évaluation sur les différences importantes entre les systèmes. Cette méthode a été utilisée avec succès à la fois avec une paire de systèmes statistique (*HTS*) et une paire de système par sélection d'unités. Les résultats que nous obtenons pour des tests de préférence AB sont clairement significatifs, alors que lorsque les phrases à vocaliser sont sélectionnées aléatoirement, les écarts de performance observés n'arrivaient pas à discriminer les systèmes.

Le reste de l'article est organisé comme suit. Dans la section 3, nous présentons les systèmes que nous utilisons dans les expériences. La section 4 décrit la méthodologie pour construire les évaluations. Enfin, la section 5 présente les expériences ainsi que les résultats.

2 Corpus de parole

Pour les besoins de cette étude, deux corpus de parole sont utilisés. Le premier corpus est extrait à l'aide d'un processus entièrement automatique présentée dans (Boeffard *et al.*, 2012), à partir d'un livre audio en français. Le locuteur masculin réalise une lecture moyennement expressive et le signal est échantillonné à $44,1$ kHz. Le corpus annoté complet contient 3339 énoncés (10h45 de parole). Pour les expériences, 1h de parole a été extraite du corps pour former le système de synthèse à base de HMM, décrit plus loin. Par la suite, ce corpus sera appelé *Audiobook*.

Le deuxième corpus est produit par un locuteur féminin en français. Il a été initialement construit pour le système de synthèse d'un serveur vocal interactif utilisé par un opérateur de télécommunications. Ses annotations ont été contrôlées manuellement. Le corpus complet contient 7h de parole enregistrées à 16 kHz. Dans la suite, ce corpus est appelé *SVI*.

3 Systèmes de synthèses de la parole

Afin d'évaluer l'efficacité de la méthode proposée, deux systèmes de synthèse de la parole sont utilisés. Le premier est basée sur *HTS* et le second est un système de Synthèse Par Corpus (*SPC*).

3.1 Synthèse par HMM

Au cours de la dernière décennie, le système *HTS* a été largement popularisé et utilisé pour de nombreuses études. Ce système de synthèse fondé sur les modèles de Markov cachés (*HMM-Based*) s'est révélé être une méthode très flexible pour produire de la parole (Tokuda *et al.*, 2002; Zen *et al.*, 2009). Cette méthode statistique repose sur la structure de modèles semi-markoviens cachés pour modéliser les coefficients Mel-généralisés (MGC), l'apériodicité, la fréquence fondamentale ($F0$) comme des flux séparés et, à l'aide d'arbres de décision, les associer à un ensemble de descripteurs (Zen *et al.*, 2009). Nous avons utilisé la version 2.3 alpha d'*HTS* avec 50 coefficients MGC, 25 coefficients de bandes apériodicité (BAP) et la $F0$.

Dans cet article, nous nous concentrons volontairement sur deux ensembles de fonctionnalités simples constituées par les phonèmes étiquetés, y compris l'étiquette du phonème en cours et les étiquettes de contexte en utilisant soit des fenêtres [-1,1], soit des fenêtres [-2,2] (un ou deux phonèmes avant et un ou deux après le phonème sélectionné). Ces configurations sont choisies en suite aux travaux de (Le Maguer *et al.*, 2013) qui ont évalué l'importance des descripteurs du jeu standard proposé par *HTS* pour le français. Il se trouve que la taille de la fenêtre phonétique utilisée a été jugée comme l'un des critères les plus pertinents, mais sans arriver à observer un écart important lors de l'évaluation perceptive. Néanmoins, en appliquant la méthodologie que nous proposons, nous allons montrer qu'il existe bien une différence significative.

À partir du corpus *Audiobook*, nous avons entraîné deux systèmes *HTS* :

— *HMM-p3* : utilise uniquement comme descripteur les étiquettes du phonème courant considéré et celles du phonème précédent et suivant ;

— *HMM-p5* : utilise les descripteurs de *HMM-p3* auxquels s'ajoute les étiquettes des deux phonèmes encadrant ceux précédemment considérés.

3.2 Synthèse par sélection d'unités

3.2.1 Système de référence

Le système de synthèse par sélection d'unités utilisé dans cette étude est celui décrit dans (Guennec & Lolive, 2014). Le coût de concaténation que nous utilisons ici comporte les trois composantes que sont les distances sur les MFCC, l'amplitude et la $F0$ entre deux unités. Pour accélérer le processus de sélection, une étape de présélection similaire à celle proposé dans (Conkie *et al.*, 2000) est employée pour filtrer les unités candidates. Les filtres utilisés agissent comme un coût cible binaire et la fonction de coût à optimiser est réduite à un coût de concaténation. Nous supposons donc que deux unités candidates passant l'étape de présélection sont équivalentes en ce qui concerne le coût cible. Les filtres suivants sont utilisées dans le système de base :

— le phonème est-il dans la dernière syllabe d'une phrase ?
— le phonème est-il dans la dernière syllabe d'un groupe syntaxique ?
— le phonème est-il dans la dernière syllabe d'un mot ?
— le phonème est-il dans une syllabe à l'intonation montante ?

3.2.2 Système de comparaison

Dans le domaine de la synthèse de la parole, la réduction de corpus est un problème général largement étudié. Comme le montre la littérature, plusieurs articles ont étudié les moyens de réduire un corpus de parole ou de texte, afin de minimiser la durée d'enregistrement ou la taille des voix manipulées. En particulier, (Lambert *et al.*, 2007) propose une évaluation de l'impact de la réduction sur la qualité d'un système de synthèse. Il y est montré qu'un corpus sélectionné aléatoirement semble produire une qualité perçue semblable à celle qui a été obtenue par l'utilisation d'un corpus construit par simple couverture des diphones. Ce point particulier est étudié en utilisant la méthodologie que nous proposons.

Le problème de la réduction de corpus peut être considérée comme un problème de couverture d'ensemble (SCP) (François & Boeffard, 2002). Celui-ci étant un problème NP-difficile, la stratégie la plus fréquente utilisée pour le résoudre repose sur des algorithmes gloutons. Compte tenu de la répartition des attributs souhaités dans les corpus linguistiques, de nombreux types d'algorithmes gloutons ont été étudiés, par exemple dans (François & Boeffard, 2002) et (Krul *et al.*, 2007). À l'aide de la relaxation lagrangienne, (Chevelu *et al.*, 2008) montre qu'un algorithme glouton par agglomération suivi d'un algorithme glouton de type *cracheur* est proche de la solution optimale.

Pour évaluer notre méthodologie, nous proposons de réduire un corpus de parole (le corpus *Complet*) en utilisant deux méthodes :

— *TTSCouv* couvre au minimum une fois chaque paire successive de phonèmes pour chaque configuration de descripteurs existante. Les descripteurs retenus sont ceux utilisé dans le moteur de synthèse décrit en 3.2.1.
— *CompAléa* est construit en complétant aléatoirement une couverture des diphonèmes réalisée sur *Complet* ($\sim$ 300 phrases) jusqu'à atteindre le même nombre de phones que dans *TTSCouv*.

En utilisant le corpus *SVI*, dont les principales statistiques sont présentées dans le tableau 1, deux systèmes de synthèse par sélection d'unités sont alors construits : ils réalisent les synthèses en utilisant respectivement *TTSCouv* et *CompAléa*, et ils sont appelés par le nom de leur corpus associé, sans risque de confusion.

Sub-corpus	*Complet*	*TTSCouv*	*CompAléa*
Durée	7h06'12	3h11'15	3h04'19
Taille en phrases	7,662	3,238	3,350
Taille en étiquettes	259,684	112,324	112,324
Nombre d'étiquettes	34 phonèmes et 2 *Non Speech sound* (NSS)		
Nombre de diphonèmes	1,242		

TABLE 1: Statistiques principales des corpus utilisés.

4 Méthodologie d'évaluation

Dans cette section, la méthode d'évaluation proposée et le corpus de test utilisé sont présentés.

4.1 Approche

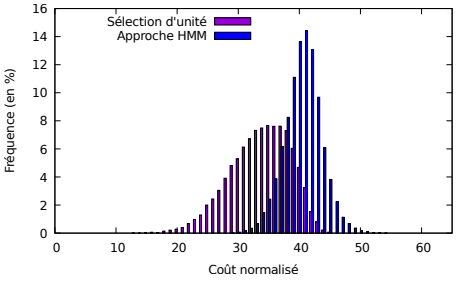
En général, l'approche classique pour les évaluations subjectives est de synthétiser un petit ensemble d'échantillons, de les proposer aux auditeurs, et de tirer des conclusions sur les systèmes à partir des résultats obtenus sur cette petite série d'échantillons. À notre avis, cette méthode fonctionne uniquement pour les systèmes qui ont une grande différence de qualité de sortie et dépend en grande partie de l'ensemble des phrases choisies. Pour révéler les différences entre deux systèmes, il faut cependant se concentrer sur les différences constatées dans les signaux de parole générés. Lorsque les évaluations se fondent sur un petit ensemble d'échantillons, les signaux les plus différents sont bien souvent absents. Par conséquent, nous proposons ce qui suit :

1. Synthétiser un grand nombre de textes provenant de domaines variés et de styles différents avec chaque système ;
2. Calculer pour chaque paire d'échantillons un coût d'alignement (par exemple une DTW – *Dynamic Time Warping* (Sakoe & Chiba, 1978)) ;
3. Sélectionner les échantillons les plus dissemblables pour évaluer les systèmes.

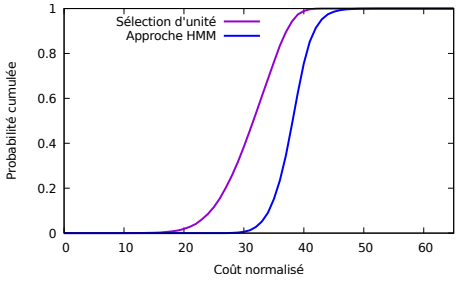
Dans le présent document, le coût d'alignement est calculé en fonction du coût de la DTW entre les vecteur de MFCC pour chaque signal, divisé par la longueur du chemin d'alignement. On obtient ainsi un coût normalisé. Cette mesure a l'avantage d'être indépendante des systèmes en cours d'évaluation mais d'autres peuvent être envisagées.

4.2 Corpus d'évaluation

Pour être indépendant des corpus de parole choisis, nous avons utilisé un corpus textuel provenant d'une source différente. Il est composé d'un ensemble de phrases extraites d'une collection de 50 livres électroniques couvrant de nombreux sujets et styles d'écriture. Les phrases qui en résultent sont ensuite filtrées pour garder celles comportant entre 30 et 60 phonèmes afin de produire des signaux d'une durée comprise approximativement entre 3 et 6 secondes (comme recommandé dans (ITU-T, 1996)). Puisque le même phonétiseur est utilisé pour les deux systèmes, les phrases avec des erreurs de phonétisation sont filtrées (généralement des phrases contenant des symboles non-standard ou des



(a) Histogrammes des valeurs des coûts.



(b) Fonctions de densité cumulée pour la mesure des coûts.

FIGURE 1: Distributions des coûts de DTW normalisés entre deux paires de systèmes évalués. Cette figure montre que la distribution a un comportement de type gaussien et qu'un nombre important d'échantillons semblent acoustiquement similaires.

noms propres). Enfin, parmi les phrases restantes, 27 030 sont extraites au hasard pour construire le corpus de test à synthétiser lors des expériences.

5 Expériences et résultats

5.1 Distribution des coûts d'alignement

La figure 1 montre la répartition des coûts de DTW normalisés pour les 27 030 phrases lorsque l'on compare *TTSCouv* et *CompAléa* (en rouge) et lorsque l'on compare *HMM-p3* et *HMM-p5* (en bleu). Considérant à la fois l'histogramme et la fonction de densité, on observe un comportement de type gaussien. La conséquence est que, lorsque des échantillons sont choisis aléatoirement, l'ensemble utilisé pour les évaluations perceptives peut contenir un nombre élevé d'échantillons acoustiquement équivalents. Ceci risque donc de lisser les résultats de l'évaluation perceptive et les systèmes peuvent être considérés comme équivalents.

Notons que les coûts pour les systèmes à base de HMM sont plus important en moyenne que ceux des systèmes par sélection d'unités. Ce résultat est prévisible puisque les deux systèmes par sélection sont construits à partir de la même voix, les signaux de sortie peuvent donc partager certains segments, ce qui n'est pas le cas pour les systèmes à base de HMM. Ainsi, malheureusement, il semble difficile de trouver un seuil universel sur le coût de DTW à partir duquel on pourrait dire que deux signaux sont significativement différents.

5.2 Évaluations perceptives

Afin d'évaluer la méthodologie proposée, nous avons mené des évaluations séparées pour les systèmes par sélection d'unités et ceux utilisant des HMM. Dans le premier cas, nous avons évalué trois méthodes d'échantillonnage. Le premier test consiste à sélectionner les échantillons de parole les plus similaires selon la mesure proposée et est fait pour vérifier que la mesure est corrélée à la perception

en termes de similarité. Le second est la méthode classique, à savoir sélectionnant de manière aléatoire un sous-ensemble d'échantillons. Enfin, le troisième test est basé sur la sélection des échantillons de parole les plus dissemblables. Dans le second cas, pour les systèmes à base de HMM, nous n'avons évalué que le sous-ensemble aléatoire d'échantillons et un sous-ensemble composé des échantillons vocaux les plus différents. Les statistiques de chaque corpus de test sont présentées dans le tableau 2. Elles montrent une différence significative entre les corpus de coûts maximaux et les autres.

Corpus de test	Nb. de phrases	Coût moyen (écart-type.)
Corpus de coûts min.	100	$15,0\ (1,6)$
Corpus aléatoire	100	$31,2\ (4,7)$
Corpus de coûts max.	100	$41,6\ (0,5)$
Corpus complet	27030	$31,2\ (4,9)$

(a) Corpus d'évaluation pour les systèmes par sélection d'unités.

Corpus de test	Nb. de phrases	Coût moyen (écart-type.)
Corpus aléatoire	100	$38,6\ (3,3)$
Corpus de coûts max.	100	$48,5\ (1,2)$
Corpus complet	27030	$34,0\ (3,0)$

(b) Corpus d'évaluation pour les systèmes par HMM.

TABLE 2: Statistiques des corpus d'évaluation.

Considérant les configurations énoncées précédemment, nous avons extrait 100 échantillons par système pour construire des tests de préférence de type AB. À chaque étape, deux signaux générés à partir de la même phrase, mais par des systèmes différents, sont présentés dans un ordre aléatoire. Il a été demandé à 10 auditeurs de choisir leur signal préféré (trois réponses sont proposées : A, B et *Indifférent*). Les résultats sont présentés dans les tableaux 3a et 3b.

Premièrement, nous pouvons observer que lors de la sélection des échantillons choisis aléatoirement, les systèmes ne sont pas différenciés et la préférence est uniformément répartie entre les trois choix possibles. Cela est vrai tant pour les systèmes par HMM que pour la sélection d'unités. En outre, la différence entre les systèmes n'est pas significative, selon un test binomial avec un intervalle de confiance à 95%. Cela peut s'expliquer par le fait que le choix aléatoire a tendance à sélectionner des échantillons contenant les événements les plus fréquents. On peut en outre supposer que, sur les événements les plus fréquents, deux systèmes proches peuvent se comporter de la même manière.

En sélectionnant les échantillons via la méthode de classement proposée dans le présent article et en ne gardant que les plus dissemblables pour l'évaluation, les résultats montrent clairement une préférence pour un système. À chaque fois, les systèmes que l'on pouvait considérer comme probablement meilleurs (*TTSCouv* et *HMM-p5*) obtiennent effectivement les meilleurs résultats. En outre, dans les deux cas, le nombre de réponses de type *Indifférent* diminue considérablement (par exemple, il est divisé par 2 pour les systèmes par sélection d'unités). Pour les deux systèmes, les résultats des tests perceptifs sont maintenant significatifs. Par conséquent, le classement proposé à permis de concentrer les tests sur un sous-ensemble d'échantillons pour lesquels les différences au niveau acoustique permettent de discriminer les systèmes évalués. Par ailleurs, on peut noter qu'aucune hypothèse n'a été faite sur la qualité de la sortie des systèmes.

Pour compléter l'évaluation de la méthode, nous avons vérifié sur les systèmes par corpus que

Système préféré	Corpus de coûts min.	Corpus aléatoire	Corpus de coûts max.
TTSCouv	27	34	52
CompAléa	27	37	32
Indifférent	46	29	16
Différence significative	Non	Non	**Oui**

(a) Résultats pour les systèmes par sélection d'unités. Trois tests AB ont été réalisées en présentant les échantillons les plus similaires, des échantillons aléatoires et les échantillons les plus différents.

Système préféré	Corpus aléatoire	Corpus de coûts max.
HMM-p3	31	26
HMM-p5	41	51
Indifférent	28	23
Différence significative	Non	**Oui**

(b) Résultats pour les systèmes à base de HMM. Deux tests AB ont été réalisées en présentant des échantillons aléatoires et les échantillons les plus différents.

TABLE 3: Résultats des tests de préférences

les échantillons les plus similaires donnent des résultats cohérents dans les tests perceptifs. Dans le tableau 3a, nous pouvons observer que, dans ce cas, un grand nombre d'échantillons sont jugés équivalents (46 votes *Indifférent*). Le reste des votes est réparti également entre *TTSCouv* et *CompAléa*. Encore une fois, la mesure appliquée ne donne aucune indication sur la qualité des échantillons. Pour conclure, ces résultats montrent clairement que sélectionner soigneusement les échantillons utilisés lors des tests perceptifs est primordial pour l'obtention de résultats significatifs.

6 Conclusion

Dans cet article, nous avons présenté une nouvelle méthode d'évaluation perceptive fondée sur un grand ensemble de test (des milliers d'échantillons) et une mesure utilisée pour classer les échantillons appariés en termes de différences acoustiques. Nous suggérons que les échantillons choisis doivent être les plus différents possibles afin d'être en mesure d'augmenter la significativité des évaluations perceptives. Cette nouvelle idée a été appliquée avec succès sur deux systèmes à base de HMM et deux systèmes par sélection d'unités, avec des voix d'apprentissage différentes (expressive avec locuteur masculin et neutre avec une locutrice). Les évaluations perceptives ont été menées pour comparer la méthode de sélection aléatoire classique à celle que nous proposons. Les résultats montrent clairement une amélioration de la significativité des résultats et une diminution des réponses de type « indifférent ».

Cette nouvelle méthodologie, qui reste simple, peut alors aider à valider efficacement des améliorations d'un système de synthèse vocale. Elle peut également être utilisée dans un procédé industriel en vue d'organiser des tests de non-régression entre les différentes versions d'un système avec un très faible coût et de repérer les phrases les plus touchées.

À l'heure actuelle, la méthode a été appliquée à des paires de systèmes, et les travaux futurs seront faits pour étendre cette méthode à un plus grand nombre de systèmes. Une approche pourrait être de réaliser les comparaisons par paires de systèmes puis de définir un rang moyen pour sélectionner les plus différents globalement. Nous prévoyons également de comparer la DTW avec d'autres distances de signaux acoustiques qui pourraient être mieux corrélées avec les évaluations perceptives.

Références

BOEFFARD O., CHARONNAT L., MAGUER S. L. & LOLIVE D. (2012). Towards fully automatic annotation of audio books for tts. In *Proc. of LREC*.

BUCHHOLZ S., LATORRE J. & YANAGISAWA K. (2013). Crowdsourced assessment of speech synthesis. *Crowdsourcing for Speech Processing*.

CHEVELU J., BARBOT N., BOËFFARD O. & DELHAY A. (2008). Comparing set-covering strategies for optimal corpus design. In *Proc. of LREC*.

CONKIE A., BEUTNAGEL M. C., SYRDAL A. K. & BROWN P. E. (2000). Preselection of candidate units in a unit selection-based text-to-speech synthesis system. In *Proc. of ICSLP*.

FRANÇOIS H. & BOEFFARD O. (2002). The greedy algorithm and its application to the construction of a continuous speech database. In *Proc. of LREC*.

GARCIA M.-N., D'ALESSANDRO C., BAILLY G., BOULA DE MAREÜIL P. & MOREL M. (2006). A joint prosody evaluation of french text-to-speech synthesis systems. In *Proc. of LREC*.

GUENNEC D. & LOLIVE D. (2014). Unit Selection Cost Function Exploration Using an A* based Text-to-Speech System. In *Proc. of TSD*.

HINTERLEITNER F., NEITZEL G., MOLLER S. & NORRENBROCK C. (2011). An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks. In *Proc. of Blizzard Challenge Workshop*.

ITU-T (1996). Recommendation : Methods for subjective determination of transmission quality.

KING S. & KARAISKOS V. (2012). The blizzard challenge 2012. In *Proc. of Blizzard Challenge workshop 2012*.

KRUL A., DAMNATI G., YVON F., BOIDIN C. & MOUDENC T. (2007). Adaptive database reduction for domain specific speech synthesis. In *Proc. of SSW6*.

LAMBERT T., BRAUNSCHWEILER N. & BUCHHOLZ S. (2007). How (not) to select your voice corpus : Random selection vs. phonologically balanced. In *Proc. of SSW6*.

LATORRE J., YANAGISAWA K., WAN V., KOLLURU B. & GALES M. J. (2014). Speech intonation for tts : Study on evaluation methodology. In *Proc. of Interspeech*.

LE MAGUER S., BARBOT N., BOËFFARD O. *et al.* (2013). Evaluation of contextual descriptors for hmm-based speech synthesis in french. In *SSW8*.

PRAHALLAD K., VADAPALLI A., KESIRAJU S., MURTHY H. A., LATA S., NAGARAJAN T., PRASANNA M., PATIL H., SAO A. K., KING S., BLACK A. W. & TOKUDA K. (2014). The blizzard challenge 2014. In *Proc. of Blizzard Challenge workshop 2014*.

SAINZ I., NAVAS E., HERNAEZ I., BONAFONTE A. & CAMPILLO F. (2014). Tts evaluation campaign with a common spanish database. In *Proc. of LREC*.

SAKOE H. & CHIBA S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*.

TOKUDA K., ZEN H. & BLACK A. W. (2002). An HMM-based speech synthesis system applied to English. In *IEEE Workshop 2002*.

VISWANATHAN M. & VISWANATHAN M. (2005). Measuring speech quality for text-to-speech systems : development and assessment of a modified mean opinion score (mos) scale. *Computer Speech & Language*.

ZEN H., TOKUDA K. & BLACK A. (2009). Statistical parametric speech synthesis. *Speech Communication*.