De l'utilisation de descripteurs issus de la linguistique computationnelle dans le cadre de la synthèse par HMM

Sébastien Le Maguer[1] Bernd Möbius[1] Ingmar Steiner[1,2] Damien Lolive[3]
(1) Computational Linguistics and Phonetics, Saarland University, Saarbrücken, Allemagne
(2) DFKI, Saarbrücken, Allemagne
(3) IRISA, Lannion, France
{slemaguer|moebius|steiner}@coli.uni-saarland.de
damien.lolive@irisa.fr

RÉSUMÉ

Durant les dernières décennies, la modélisation acoustique effectuée par les systèmes de synthèse de parole paramétrique a fait l'objet d'une attention particulière. Toutefois, dans la plupart des systèmes connus, l'ensemble des descripteurs linguistiques utilisés pour représenter le texte reste identique. Plus specifiquement, la modélisation de la prosodie reste guidée par des descripteurs de bas niveau comme l'information d'accentuation de la syllabe ou bien l'étiquette grammaticale du mot. Dans cet article, nous proposons d'intégrer des informations basées sur la prédictibilité d'un évènement (la syllabe ou le mot). Plusieurs études indiquent une corrélation forte entre cette mesure, fortement présente dans la linguistique computationnelle, et certaines spécificités lors de la production humaine de la parole. Notre hypothèse est donc que l'ajout de ces descripteurs améliore la modélisation de la prosodie. Cet article se focalise sur une analyse objective de l'apport de ces descripteurs sur la synthèse HMM pour la langue anglaise et française.

ABSTRACT

Toward the use of information density based descriptive features in HMM based speech synthesis

Over the last decades, acoustic modeling for speech synthesis has been improved significantly. However, in most systems, the descriptive feature set used to represent annotated text has been the same for many years. Specifically, the prosody models in most systems are based on low level information such as syllable stress or word part-of-speech tags. In this paper, we propose to enrich the descriptive feature set by adding a linguistic measure computed from the predictability of an event, such as the occurrence of a syllable or word. By adding such descriptive features, we assume that we will improve prosody modeling. This new feature set is then used to train prosody models for speech synthesis. This paper focuses on an objective analysis of the influence of these descriptive features on the synthesis achieved in English and French.

MOTS-CLÉS : Synthèse de la parole paramétrique, densité d'information, descripteurs linguistiques.

KEYWORDS: Parametric speech synthesis, information density, descriptive features.

1 Introduction

Durant ces dernières années, la popularité de la synthèse paramétrique a pris de l'ampleur au point de devenir l'une des méthodologies standards de la synthèse text-to-speech (TTS). De la synthèse par hidden Markov model (HMM) (Zen & Toda, 2005) aux réseaux de neurones profonds (Zen *et al.*, 2013), l'effort de recherche s'est principalement focalisé sur la modélisation acoustique.

Toutefois, l'ensemble de ces systèmes se basent sur le même ensemble de descripteurs linguistiques et prosodiques pour prédire les paramètres acoustiques. Ainsi, la majorité des systèmes basés sur HMM utilisent un jeu de descripteurs dérivés du jeu proposé dans Tokuda *et al.* (2002). Peu de travaux se focalisent sur l'intégration de nouveaux descripteurs linguistiques et prosodiques. Parmi celles ci, nous pouvons citer l'utilisation de « word embeddings » Wang *et al.* (2015) ou bien l'utilisation d'informations syntaxiques enrichies (Obin *et al.*, 2010).

Dans Le Maguer *et al.* (2016), nous avons proposé d'intégrer de nouveaux descripteurs, issus du domaine de la « densité d'information » (Information Density). Les résultats montrent que l'utilisation de tels descripteurs améliore la synthèse effectuée. Ces descripteurs sont basés sur l'évaluation de l'imprédictibilité d'un évènement ; notion répandue en linguistique computationnelle. Ils sont obtenus en utilisant un modèle de langue ce qui procure plusieurs avantages. Tout d'abord, ils sont simples à obtenir à partir d'un texte et peuvent être utilisés pour aboutir à des descripteurs de plus haute abstraction linguistique. Ils peuvent également être utilisés pour la synthèse classique ou bien la synthèse incrémentale (Baumann & Schlangen, 2012; Pouget *et al.*, 2015).

Dans le présent article, nous proposons d'appliquer cette méthodologie pour une synthèse HMM pour le français. Nous proposons également d'analyser l'impact de ces descripteurs sur les modèles appris et de comparer cet impact aux résultats obtenus pour l'anglais. Cette comparaison est rendue possible par le fait que le jeu de descripteurs que nous avons utilisé pour les deux langues est similaire.

Ainsi, cet article est organisé de la manière suivante : la section 2 présente et justifie l'utilisation de tels descripteurs. La section 3 décrit le protocole d'évaluation mis en place pour analyser l'influence de ces descripteurs sur la synthèse. La dernière section (4) présente les résultats de l'évaluation objective pour le français et l'anglais.

2 Descripteurs linguistiques basés sur la densité d'information

Reposant sur la théorie de l'information proposée par Shannon (1948), Hale (2001) introduit le concept d'imprédictibilité (*surprisal*), associé à la densité d'information, au sein du domaine de la linguistique computationnelle. Il repose sur une modélisation n-gramme et est défini par l'équation suivante :

$$Surp(U_i) = -log_2(P(U_i|U_{i-1}..U_{i-1-N})) \quad (1)$$

où U_i correspond à l'unité analysée et $U_{i-1}..U_{i-1-N}$ sont les N unités précédentes. N est le paramètre du modèle et doit être défini.

L'utilisation d'un tel concept repose sur le résultat d'études issues du domaine de la psycho-linguistique. En effet, la prédictibilité d'un mot est fortement corrélée avec l'effort nécessaire pour prononcer ce mot (Smith & Levy, 2013). Une corrélation analogue a été mise en avant avec la prédictibilité d'une syllabe (Jaeger, 2010). Ainsi, notre hypothèse est que l'introduction de descrip-

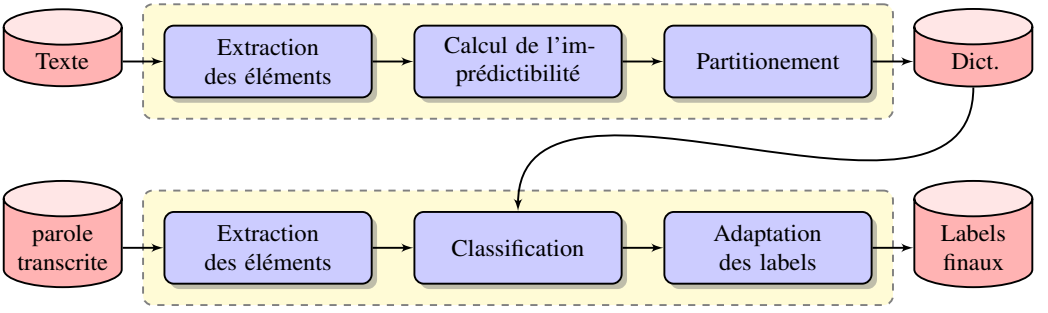

FIGURE 1 – Génération des descripteurs basés sur la densité d'information. À partir du corpus de texte, nous souhaitons extraire un dictionnaire qui associe un cluster d'imprédictibilité à chaque élément (syllabe ou mot). Tout d'abord, nous extrayons les éléments en utilisant des outils TALN et nous calculons leur imprédictibilité via (1). Ensuite, nous partitionnons l'espace obtenu afin de déterminer une échelle allant du plus prédictible (valeur à 0) au moins prédictible. L'étape finale consiste à associer à chaque élément une partition en se basant sur l'imprédictibilité de cet élément. À partir du corpus de parole, nous extrayons le même type d'élément et, grâce au dictionnaire, déterminons la partition associée. Enfin, nous adaptons les fichiers de labels afin de les rendre compatibles avec le système de synthèse.

teurs linguistiques basés sur la prédictibilité devrait améliorer la synthèse et plus spécifiquement la modélisation de la prosodie. En effet, concernant la synthèse par HMM, la modélisation du spectre est davantage contrôlée par les informations phonologiques (Le Maguer *et al.*, 2013).

Toutefois, l'utilisation de l'imprédictibilité a deux problèmes principaux. Le premier est que, pour obtenir des statistiques valides, il nous faut disposer d'un corpus textuel beaucoup plus large que les corpus de paroles généralement utilisés. Le second est que la prédictibilité est un descripteur à valeurs continues alors que les systèmes TTS standards reposent sur l'utilisation de descripteurs à valeurs discrètes.

2.1 Processus général

Afin de palier ces problèmes, nous proposons l'utilisation du processus décrit dans la figure 1.

Tout d'abord, nous avons décidé d'utiliser deux corpora : un corpus de parole et un corpus de textes beaucoup plus large que le précédent. À partir du corpus de texte, nous déterminons l'imprédictibilité des évènements analysés et nous générons un dictionnaire de classes d'imprédictibilité. Ces classes sont définies en utilisant un algorithme de partitionnement, dans notre cas celui des k-moyennes. Ceci nous permet d'obtenir une approximation discrète des valeurs continues.

2.2 Descripteurs associés à la syllabe

Puisque nous utilisons deux corpora, pouvant être obtenus par différents outils TALN, nous proposons d'utiliser une représentation reposant sur l'alphabet phonétique international (International Phonetic Alphabet (IPA)) pour les phonèmes constituant chaque syllabe. Il est possible d'ajouter des infor-

mations plus spécifiques. Toutefois, dans cette étude, nous utilisons uniquement une représentation IPA.

2.3 Descripteurs associés au mot

Afin de représenter les mots, nous devons rester le plus possible proche du texte. Néanmoins, pour obtenir une représentation adaptée, il est nécessaire de nettoyer ce texte. Nous avons procédé de la manière suivante :

— tous les signes de ponctuation sont supprimés ;
— un indicateur typographique (le symbole #) est inséré à la fin de chaque paragraphe ;
— tous les mots sont convertis en minuscule.

Les indicateurs sont traités comme des mots à part entière. Ils ont été insérés à la fin de chaque paragraphe car nous faisons l'hypothèse qu'un paragraphe est conceptuellement homogène (le paragraphe ne traite que d'un seul « sujet »). L'apprentissage de modèles de parole est généralement basé sur des énoncés globalement courts. Ainsi, l'utilisation d'indicateurs est plus cohérente que l'utilisation d'une valeur non-définie en début de chaque énoncé. En effet, ajouter un indicateur permet aux arbres de décision de prendre en compte des partitions qui auraient été autrement ignorées.

3 Protocole expérimental

3.1 Corpus anglais

Le corpus anglais est issu du challenge Blizzard 2013 (King & Karaiskos, 2013). Il est composé de 83 livres audio lus par un locuteur féminin. De ce jeu de données, 2 corpora sont extraits : le *corpus de textes* et le *corpus de parole*. Le *corpus de textes* est composé de l'ensemble du corpus anglais excepté le livre « Black Beauty ». Cela correspond à 82 livres, 2 298 055 syllabes et 1 973 368 mots. Le *corpus de parole* est composé d'environ une heure de parole (∼470 énoncés) extraits de « Black Beauty » soit 13 522 syllabes et 7038 mots. Pour chaque corpus, les frontières de syllabes ont été obtenues grâce au système MaryTTS (Schröder & Trouvain, 2003) (version 5.2).

3.2 Corpus français

Le corpus français est constitué du roman « À la recherche du temps perdu » de Marcel Proust. De manière analogue au corpus anglais, 2 corpora sont extraits de ce jeu de données. Le *corpus de textes* est composé de l'ensemble du corpus français excepté le tome « Albertine disparue ». Cela correspond à 6 tomes, 1 806 672 syllabes et 1 005 492 mots. Le *corpus de parole* est composé d'environ une heure de parole (∼520 énoncés) extraits de « Albertine disparue » soit 15 088 syllabes et 10 051 mots. Pour chaque corpus, les frontières de syllabes ont été obtenues par un système par règles.

3.3 Analyse de l'imprédictibilité sur les corpus

Pour déterminer l'imprédictibilité, nous utilisons des trigrammes de mots et des trigrammes de syllabes.

Pour le corpus de parole anglais, l'ensemble des trigrammes des mots est présent dans le corpus de textes. En revanche, pour le corpus français, 33 % sont absents. Pour les trigrammes de mots, environ 40 % des instances du corpus de parole sont absentes du corpus de textes pour l'anglais ; contre 80 % pour le corpus français. En considérant les unités qui sont présentes dans les deux corpus, pour chaque langue, la figure 2 illustre la distribution de la prédictibilité des syllabes et mots distincts.

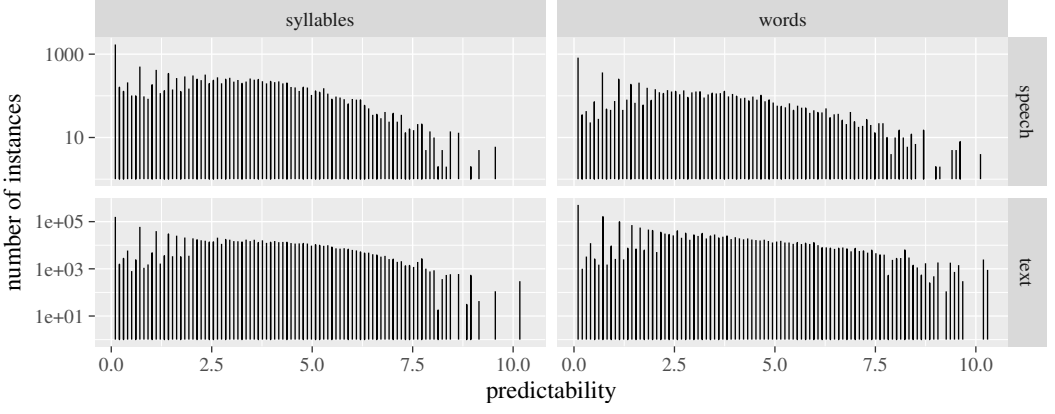

FIGURE 2 – Distribution de l'imprédictibilité des syllabes et des mots pour les corpus de textes et de parole

L'ensemble des distributions respectent le même schéma. Une imprédictibilité de 0 indique que l'évènement est qualifié intégralement par son contexte. En effet, les séquences $U_{i-1}...U_{i-N}$ et $U_i...U_{i-N}$ n'apparaissant qu'une seule fois dans le corpus ; la détection de $U_{i-1}...U_{i-N}$ implique forcément l'apparition de U_i.

Ensuite, nous avons les évènements fréquents avec une valeur d'imprédictibilité faible. Ces évènements correspondent aux schémas linguistiques courants tels que « one of those » pour l'anglais. L'utilisation de trigrammes, comparativement à une modélisation n-gramme plus large, a accru le nombre de ces évènements. Enfin, le nombre d'occurrences des éléments décroits au fur et à mesure que leur imprédictibilité augmente.

Le dernier paramètre du modèle correspond au nombre de partitions utilisées lors de la phase de partitionnement. Pour les expériences effectuées, nous avons utilisé 9 partitions ce qui correspond à la configuration par défaut de l'outil utilisé. En conséquence, l'imprédictibilité d'un évènement est définie sur une échelle 0 à 8, avec 0 indiquant une unité intégralement définie par son contexte, et 8 une unité dont les occurrences sont les plus inattendues.

3.4 Analyse des descripteurs linguistiques

Afin d'analyser l'influence des descripteurs proposés, nous avons défini trois jeux de descripteurs *baseline*, *pred_syllable*, and *pred_all*.

Le jeu *baseline* correspond au jeu proposé par Tokuda *et al.* (2002) pour l'anglais et par Le Maguer *et al.* (2013) pour le français. Les jeux sont similaires pour les deux langues mais obtenus via des outils logiciels différents. Dans les deux cas, l'information de stress associée à la syllabe a été écartée et il n'y a pas d'équivalent d'étiquette ToBI pour le jeu français.

Les deux autres configurations intègrent, respectivement, l'information d'imprédictibilité de la syllable puis cette même information pour la syllabe et le mot.

3.5 Système de synthèse

Afin d'analyser les descripteurs proposés, le système de synthèse utilisé est la configuration standard du système HMM based speech synthesis system (HTS) (Zen & Toda, 2005) (version 2.3) avec le vocodeur STRAIGHT (Kawahara *et al.*, 1999).

4 Analyse des modèles obtenus

Lors de notre première analyse pour l'anglais, présenté dans Le Maguer *et al.* (2016), nous avons constaté que l'évaluation par distance n'indiquait pas d'amélioration de la similarité. En revanche les évaluations subjectives montrent une amélioration perceptible de la synthèse.

Ainsi, afin d'analyser l'influence de ces descripteurs sur la synthèse, nous allons nous focaliser sur l'évolution de la structure des arbres de décision suivant les différentes configurations.

Pour les arbres de décisions du système HTS, les nœuds correspondent à des propriétés associées aux descripteurs linguistiques et les feuilles correspondent aux modèles eux mêmes. En conséquence, en considérant le corpus d'apprentissage et un arbre de décision, il est possible de déterminer l'importance accordée par le système à un descripteur. Enfin, nous avons groupé les descripteurs par catégorie afin d'aboutir à un analyse plus globale.

Dans notre cas, nous considérons les catégories suivantes :

— p$\{1, 3, 5\}$ pour la taille de la fenêtre à l'horizon du phonème (monophone, triphone, quinphone) ;
— $\{syl, word, phrase\}$-position pour les informations de position (e.g. position de la syllabe courante dans le mot courant) correspondant au niveau linguistique donné ;
— $\{syl, word, phrase\}$-prosody pour les descripteurs associés à la prosodie (accentuation de la syllabe, part-of-speech (POS) associé au mot) ;
— énoncé pour l'ensemble des informations de comptage global (nombre total de syllabes, de mots et de groupes prosodiques).

Nous y ajoutons deux catégories, $\{syl, word\}$-predictability, qui correspondent aux descripteurs proposés.

	Anglais			Français		
catégories	**baseline**	**pred_syl.**	**pred_all**	**baseline**	**pred_syl.**	**pred_all**
p1	1648	1615	1594	3202	3135	3152
p3	2174	2037	1175	2495	1170	1247
p5	0	0	694	0	892	897
syl-position	**5799**	**5652**	**4056**	**13 034**	**9405**	**9820**
syl-prosody	98	128	183	54	21	24
syl-predictability	0	1657	**4163**	0	**10 773**	**9884**
word-position	**7836**	**6787**	**4202**	**13 747**	**9235**	**10 406**
word-prosody	2928	2817	2188	**18 496**	**13 605**	**14 438**
word-predictability	0	0	**7834**	0	0	4548
phrase-position	1184	1573	802	2833	2878	3147
phrase-prosody	**8723**	**8323**	**5892**	0	0	0
utterance	**7260**	**7429**	**6799**	**16 421**	**13 922**	**15 139**

TABLE 1 – Analyse de l'évolution de l'arbre de décision du F0 associé à l'état central du HMM en utilisant les labels du corpus de parole. Chaque label est passé à travers l'arbre. À chaque nœud de l'arbre est associée une catégorie et chaque fois qu'un nœud est atteint le compteur de la catégorie correspondante est mis à jour. Les catégories les plus utilisées ont été surlignées.

Enfin, nous avons appliqué la normalisation suivante :

$$\text{Freq\_occupation (noeud)} = \frac{\text{Freq\_occupation (noeud)}}{\text{Nb\_classes(descripteur(noeud))}} \tag{2}$$

Cela permet d'éviter d'accorder une importance injustifiée à une catégorie simplement parce que le nombre de propriétés associées à celle-ci est plus importante que pour les autres catégories.

En considérant les différents arbres, l'information de prédictibilité a un impact limité sur la modélisation du spectre et de l'apériodicité. En effet, pour ces arbres, les informations principalement utilisées correspondent aux étiquettes des phonèmes. En revanche, Les arbres de décision associés à la durée et au F0 sont impactés par cette catégorie de descripteurs. De plus, les arbres associés à ces deux paramètres ont une structure similaire.

En conséquence, nous analysons uniquement les résultats obtenus pour l'arbre de décision de l'état central du HMM pour le F0. Ces résultats sont présentés dans le tableau 1.

Tout d'abord, en comparant les résultats pour le français et pour l'anglais, on peut remarquer une différence significative du nombre d'utilisation de nœuds (maximum 8723 pour l'anglais, 18496 pour le français). Ceci s'explique par la différence de durée entre les corpus mais également par un débit différent des locuteurs. En effet, le corpus anglais contient environ 20 % de segments en moins.

En se focalisant sur la configuration par défaut (*baseline*), la principale différence entre les deux langues repose sur l'utilisation du tag grammatical et les informations de position à l'horizon du groupe de souffle. Ceci s'explique par le fait que les questions (et valeurs possibles des descripteurs) associées au tag grammatical sont plus précises et nombreuses en français. En effet, pour l'anglais,

il n'y aucune distinction entre les tags « signifiants ». En revanche, le jeu de descripteurs français distingue les verbes, noms,...

Étonnamment, dans les deux cas, l'information d'accentuation n'est utilisée que pour affiner le modèle ; elle n'est pas considérée par HTS, pour ces jeux de données, comme une propriété fondamentale du F0. Ceci peut s'expliquer par le fait que l'information d'accentuation peut être capturée par l'information de position associée à la syllabe.

Prendre en compte l'imprédictibilité à l'horizon de la syllabe n'impacte pas fondamentalement les modèles pour le corpus anglais. En revanche, pour le corpus français, ce descripteur devient l'un des plus utilisé. En comparant la répartition d'occupation de nœuds pour le français, nous constatons que les informations de position à l'horizon de la syllabe et du mot sont fortement impactés par l'introduction de l'imprédictibilité à l'horizon de la syllabe. Cet impact est également visible pour l'anglais mais de manière plus limité. L'imprédictibilité étant déterminée sur un contexte de trois syllabes, il n'est pas étonnant d'aboutir à un tel résultat. En effet, l'information du nombre de syllabes contenues dans un mot est implicitement encodé dans l'information d'imprédictibilité de part l'introduction de ce contexte.

Enfin, prendre en compte l'imprédictibilité à l'horizon du mot pour le français ne fait que moyenner l'utilisation des autres catégories. En revanche, pour l'anglais, l'ajout de cette information aboutit à accorder plus d'importance à l'imprédictibilité à l'horizon de la syllabe également. Ainsi, d'après les modèles, la prosodie du locuteur français utilisé pourrait se situer d'avantage à l'horizon de la syllabe. Ceci serait cohérent avec le constat effectué précédemment concernant l'information de prédictibilité à l'horizon de la syllabe.

Ainsi, le système HTS considère l'information de prédictibilité comme information importante pour l'ensemble des deux langues. Toutefois, l'accent semble plutôt mis sur l'horizon de la syllabe pour le corpus français et sur l'horizon du mot pour le corpus anglais.

5 Conclusion

Dans cet article nous avons intégré un nouveau type de descripteurs linguistiques pour l'anglais et le français. Nous avons comparé l'influence de ces descripteurs sur la modélisation effectuée par le système HTS. Nous avons pu constater que cette information était considérée comme importante pour les deux corpus par le système HTS.

Toutefois, il reste difficile d'évaluer objectivement et subjectivement l'apport de tels descripteurs. En effet, les méthodes d'évaluation subjective permettant d'évaluer la qualité de modélisation des paramètres prosodiques, utilisés dans la synthèse de la parole, indépendamment des paramètres spectraux restent à déterminer.

6 Remerciements

La recherche présentée ici à été financé par la German Research Foundation (DFG) via le projet SFB 1102 « Information Density and Linguistic Encoding » à l'université de la Sarre. Nous souhaitons enfin remercier Marina Oberwegner pour la correction manuelle de la segmentation.

Références

BAUMANN T. & SCHLANGEN D. (2012). INPRO_iSS : a component for just-in-time incremental speech synthesis. In *Proceedings of the ACL 2012 System Demonstrations*, p. 103–108 : Association for Computational Linguistics.

HALE J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, p. 1–8 : Association for Computational Linguistics.

JAEGER T. F. (2010). Redundancy and reduction : speakers manage syntactic information density. *Cognitive Psychology*, **61**, 23–62.

KAWAHARA H., MASUDA-KATSUSE I. & DE CHEVEIGNÉ A. (1999). Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction : Possible role of a repetitive structure in sounds. **27**, 187–207.

KING S. & KARAISKOS V. (2013). The Blizzard Challenge 2013.

LE MAGUER S., BARBOT N. & BOEFFARD O. (2013). Evaluation of contextual descriptors for hmm-based speech synthesis in french. In *Proceedings of the Speech Synthesis Workshop (SSW)*, Barcelona (Spain).

LE MAGUER S., MÖBIUS B. & STEINER I. (2016). Toward the use of information density based descriptive features in hmm based speech synthesis. In *Proceedings of Speech Prosody*. to appear.

OBIN N., RODET X. & LACHERET A. (2010). Hmm-based prosodic structure model using rich linguistic context. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, p. 1133–1136.

POUGET M., HUEBER T., BAILLY G. & BAUMANN T. (2015). HMM training strategy for incremental speech synthesis. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*.

SCHRÖDER M. & TROUVAIN J. (2003). The German text-to-speech synthesis system MARY : A tool for research, development and teaching. *International Journal of Speech Technology*, **6**, 365–377.

SHANNON C. (1948). *A mathematical theory of distribution*, volume 27.

SMITH N. J. & LEVY R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, **128**(3), 302–319.

TOKUDA K., ZEN H. & BLACK A. W. (2002). An HMM-based speech synthesis system applied to English. In *Proceedings of the Speech Synthesis Workshop (SSW)*.

WANG P., QIAN Y., SOONG F. K., HE L. & ZHAO H. (2015). Word embedding for recurrent neural network based TTS synthesis. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, p. 4879–4883.

ZEN H., SENIOR A. & SCHUSTER M. (2013). Statistical parametric speech synthesis using deep neural networks. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, p. 7962–7966.

ZEN H. & TODA T. (2005). An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In *Proceedings of the European Conference on Speech Communicationand Technology (Eurospeech)*.