Une Description Probabiliste de la Communication Parlée entre Homme et Machine

Titre anglais : A Probabilistic Framework for Man-Machine Communication.

Auteur : Olivier Pietquin
Adresse : Philips GmbH Research Labs. HDS Dept., Weisshausstrasse 2, D-52066 Aachen, Allemagne
Email : olivier.pietquin@philips.com
Tel : +49 241 6003 643
Fax : +49 241 6003 518

Résumé : Cet article présente une tentative de formalisation de la communication parlèe entre homme et machine dans le cadre des systèmes de dialogues vocaux. Cette formalisation se base sur une description probabiliste de l'interaction et du traitement de l'information contenue dans le signal de parole par les différents modules qui composent un système de dialogue homme-machine. Une série d'applications possibles de ce cadre théorique à la conception de systèmes de dialogues est enfin proposée.

Abstract : This paper presents a formalism for the description of man-machine spoken communication in the framework of spoken dialogue systems. This formalisation is based on a probabilistic description of the interaction and of the information processing occurring in each module composing a spoken dialogue system. Eventually, some possible applications of this theoretic framework to the problem of dialogue systems' design are proposed.

Mots Clés : Dialogue Homme-Machine, Interaction Vocale, Traitement du Langage Naturel

Logiciel : MS Word 2000, GhostScript 7.04, GSView 4.3

Formes de la soumission : Article Court

Thèmes : Interaction vocale, Formalismes et IHM

Une Description Probabiliste de la Communication Parlée entre Homme et Machine

Olivier Pietquin

Philips GmbH Research Labs.
Home Dialogue Systems Dept.
Weisshausstrasse, 2
D-52066 Aachen, Allemagne
olivier.pietquin@philips.com

Faculté Polytechnique de Mons
TCTS Lab.
Rue de Houdain, 9
B-7000 Mons, Belgique
pietquin@tcts.fpms.ac.be

RESUME
Cet article présente une tentative de formalisation de la communication parlèe entre homme et machine dans le cadre des systèmes de dialogues vocaux. Cette formalisation se base sur une description probabiliste de l'interaction et du traitement de l'information contenue dans le signal de parole par les différents modules qui composent un système de dialogue homme-machine. Une série d'applications possibles de ce cadre théorique à la conception de systèmes de dialogues est enfin proposée.

MOTS CLES : Dialogue Homme-Machine, Interaction Vocale, Traitement du Langage Naturel

ABSTRACT
This paper presents a formalism for the description of man-machine spoken communication in the framework of spoken dialogue systems. This formalisation is based on a probabilistic description of the interaction and of the information processing occurring in each module composing a spoken dialogue system. Eventually, some possible applications of this theoretic framework to the problem of dialogue systems' design are proposed.

CATEGORIES AND SUBJECT DESCRIPTORS: G.3 [Probability and Statistics]: Probabilistic Algorithms, Markov Process; H.5.1 [Multimedia Information Systems]: Audio input/output, Evaluation/Methodology.

GENERAL TERMS: Theory, Design

KEYWORDS : Spoken Dialogue Systems, Speech Processing, Natural Language Processing.

INTRODUCTION
Durant les dernières décennies, les techniques de traitement de la parole (comme la reconnaissance et la synthèse de parole) et de traitement automatique du langage naturel (comme la compréhension et la génération de textes) ont évolué de manière importante. Il est aujourd'hui techniquement possible de réaliser des interfaces homme-machine utilisant la voix comme moyen de communication principal. Néanmoins, la percée des interfaces vocales est moins spectaculaire que celle qui nous était prédite voici quelques années. La raison en est principalement que la réalisation d'une interface vocale ne consiste pas simplement en l'adjonction de technologies les unes aux autres. Il est, en effet, nécessaire de réaliser une gestion intelligente de l'interaction en tenant compte des performances objectives des différentes technologies mises en œuvre (qui, étant donné leur caractère probabiliste, ne sont pas fiables à cent pourcents), et des performances désirées en matière d'ergonomie et d'accomplissement des tâches de l'interface.

Afin de réaliser le développement d'interfaces vocales de manière plus systématique et d'optimiser objectivement le résultat obtenu, il est bon de formaliser le problème dans un cadre plus mathématique. C'est ce que nous nous proposons de faire dans cet article.

DESCRIPTION DE LA COMMUNICATION PARLEE ENTRE HOMME ET MACHINE
Le traitement du langage est un vaste problème qui a donné naissance à plusieurs champs de recherche. Le développement d'interfaces vocales fait intervenir un certain nombre d'entre eux comme la reconnaissance vocale (ASR[1]) [2], la compréhension de langage naturel (NLU[2]) [1], la génération automatique de texte (NLG[3]) [6], la synthèse vocale (TTS[4]) [3] et la gestion de dialogue comme le montre le schéma de la Figure 1.

Cette figure indique aussi le parcours de l'information au sein du système de dialogue ainsi que ses diverses

[1] Automatic Speech Recognition
[2] Natural Language Understanding
[3] Natural Language Generation
[4] Text-To-Speech

transformations. Afin de décrire ce parcours dans des termes intégrables dans une description formelle, nous allons tout d'abord définir un dialogue homme-machine comme un processus séquentiel composé de cycles que l'ont peut observer en des temps discrets. Le laps de temps entre deux observations peut être de longueur variable et il compose un *tour*. A chaque tour t, le système de gestion de dialogue accomplit une action $\boldsymbol{a_t}$ qu'il choisit en accord avec sa stratégie interne (le développement de cette stratégie est généralement le processus le plus critique dans la conception d'une interface vocale) et en fonction de l'historique de l'interaction (représenté par la succession de ses états internes $\{\boldsymbol{s_t}\}_{t=0,...t}$).

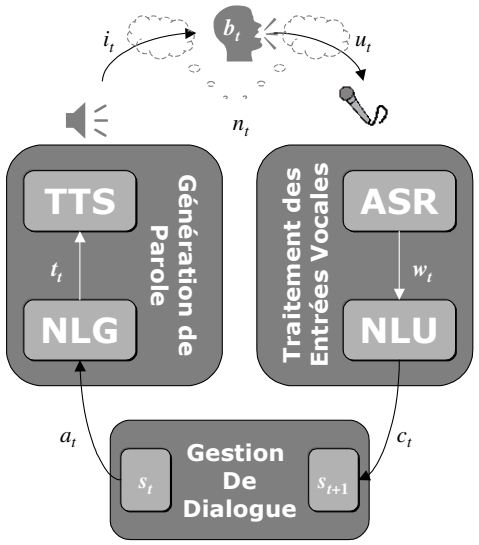

Figure 1 : Description de la communication parlée entre homme et machine

Typiquement, une action peut être une question posée à l'utilisateur, une demande de confirmation, la présentation d'une information demandée par l'utilisateur, l'ouverture ou la fermeture du dialogue etc. Cette action est ensuite transformée en une séquence de mots, un texte $\boldsymbol{t_t}$, exprimant les concepts qu'elle intègre par le sous-système de génération automatique de texte (NLG). Ce texte est alors utilisé par le module de synthèse de parole (TTS) pour créer le signal de parole $\boldsymbol{i_t}$, destiné à l'utilisateur. Celui-ci, en fonction de sa compréhension de i_t et du but $\boldsymbol{b_t}$ qu'il poursuit en utilisant l'interface, produit une réponse sous forme d'un signal de parole $\boldsymbol{u_t}$ qui peut être modifié par un bruit extérieur $\boldsymbol{n_t}$ avant d'être capturé par un microphone. Le signal capturé est ensuite traité par le module de reconnaissance de parole (ASR) pour transformer u_t en une séquence de mots $\boldsymbol{w_t}$ qui est elle-même utilisée par le système de compréhension de parole (NLU) afin d'en extraire la signification et de la formuler comme une séquence de concepts $\boldsymbol{c_t}$. C'est cette dernière séquence qui est finalement utilisée par le module de gestion de dialogue pour mettre à jour son état interne et le processus peut alors recommencer.

Propriété de Markov

Un système de dialogue homme-machine possède la propriété de Markov si la décision de son module de gestion de dialogue d'accomplir l'action a_t au tour t et son état s_{t+1} au tour suivant ne dépendent que de son état s_t au tour t et pas des états précédents :

$$P(a_t, s_{t+1} \mid s_t, s_{t-1}, \ldots, s_0) = P(a_t, s_{t+1} \mid s_t)$$

Ceci ne signifie pas que le système ne doit pas tenir compte de l'historique de l'interaction pour prendre une décision quant à l'action à accomplir afin de posséder la propriété de Markov mais bien que l'état du dialogue doit contenir en lui suffisamment d'information sur l'historique pour que cette condition soit remplie. Par la suite, nous considèrerons que le système étudié possède cette propriété.

DESCRIPTION PROBABILISTE

Probabilistiquement parlant, on peut décrire le fonctionnement du système par la probabilité d'apparition conjointe des différents signaux $\{a_t, t_t, i_t, b_t, u_t, w_t, c_t, s_{t+1}\}$ étant donnés les signaux $\{s_t, n_t\}$ (ceci suppose que le système possède la propriété de Markov):

$$P(a_t, t_t, i_t, b_t, u_t, w_t, c_t, s_{t+1} \mid s_t, n_t)$$

Cette probabilité, peut être factorisée comme suit :

$$\overbrace{P(a_t \mid s_t, n_t)}^{\text{Gestion de Dialogue}} \cdot \overbrace{P(s_{t+1} \mid w_t, c_t, u_t, b_t, i_t, t_t, a_t, s_t, n_t)}^{\text{Mise a Jour de l'Etat}} \cdot \overbrace{P(b_t \mid t_t, i_t, a_t, s_t, n_t)}^{\text{Modification de but}}$$
$$\underbrace{P(t_t, i_t \mid a_t, s_t, n_t)}_{\text{Génération de Parole}} \cdot \underbrace{P(w_t, c_t \mid u_t, b_t, t_t, i_t, a_t, s_t, n_t)}_{\text{Traitement des Entrées Vocales}} \cdot \underbrace{P(u_t \mid b_t, t_t, i_t, a_t, s_t, n_t)}_{\text{Utilisateur}}$$

Plusieurs hypothèses peuvent être émises afin de simplifier cette expression :

- La mise à jour de l'état est incrémentale et ne tient compte que du supplément d'information contenu dans c_t :

$$P(s_{t+1} \mid w_t, c_t, u_t, b_t, i_t, t_t, a_t, s_t, n_t) = P(s_{t+1} \mid c_t, s_t)$$

- La gestion de dialogue ne tient pas compte du bruit (il arrive parfois que le système de reconnaissance vocale s'adapte au bruit):

$$P(a_t \mid s_t, n_t) = P(a_t \mid s_t)$$

- Le processus de génération de parole ne tient pas compte du bruit:

$$P(t_t, i_t \mid a_t, s_t, n_t) = P(t_t, i_t \mid a_t, s_t)$$

- Si l'utilisateur peut changer de but au cours de l'interaction, les seuls facteurs pouvant l'influencer sont l'état du système (l'historique) et le signal de parole qu'il reçoit du système puisqu'il ne connaît pas a_t ou t_t :

$$P(b_t \mid t_t, i_t, a_t, s_t, n_t) = P(b_t \mid i_t\, s_t)$$

- De la même manière, la réponse de l'utilisateur ne peut pas dépendre de la véritable action a_t ou du texte généré t_t :

$$P(u_t \mid b_t, t_t, i_t, a_t, s_t, n_t) = P(u_t \mid b_t, i_t, s_t, n_t)$$

- Le procédé de traitement des entrées vocales ne dépend pas du but de l'utilisateur, du texte généré ou de la parole générée par le système :

$$P(w_t, c_t \mid u_t, b_t, t_t, i_t, a_t, s_t, n_t) = P(w_t, c_t \mid u_t, a_t, s_t, n_t)$$

Après simplification, on obtient :

$$\overbrace{P(a_t \mid s_t)}^{\text{Gestion de Dialogue}} \cdot \overbrace{P(s_{t+1} \mid c_t, s_t)}^{\text{Mise a Jour de l'Etat}} \cdot \overbrace{P(b_t \mid i_t, s_t)}^{\text{Modification de but}}$$
$$\underbrace{P(t_t, i_t \mid a_t, s_t)}_{\text{Génération de Parole}} \cdot \underbrace{P(w_t, c_t \mid u_t, a_t, s_t, n_t)}_{\text{Traitement des Entrées Vocales}} \cdot \underbrace{P(u_t \mid b_t, i_t, s_t, n_t)}_{\text{Utilisateur}}$$

Traitement des Entrées Vocales

Le traitement des entrées vocales est représenté par le terme $P(w_t, c_t \mid u_t, a_t, s_t, n_t)$. Celui-ci peut encore être factorisé comme suit :

$$\overbrace{P(c_t, w_t \mid u_t, a_t, s_t, n_t)}^{\text{Traitement des Entrées Vocales}} = \overbrace{P(c_t \mid w_t, u_t, a_t, s_t, n_t)}^{\text{NLU}} \cdot \underbrace{P(w_t \mid u_t, a_t, s_t, n_t)}_{\text{ASR}}$$

En ce qui concerne la compréhension de parole, le terme peut encore une fois être simplifié puisque le procédé ne prend en compte que les mots w_t résultants de la reconnaissance vocale et pas le véritable signal u_t prononcé par l'utilisateur ni le bruit n_t qui peut avoir affecté le résultat de reconnaissance vocale :

$$P(c_t \mid w_t, u_t, a_t, s_t, n_t) = P(c_t \mid w_t, a_t, s_t)$$

Le terme associé à la reconnaissance vocale ne peut plus être simplifié. Par contre, il peut être transformé grâce à la loi de Bayes et permettre de retrouver l'équation habituelle à la base de la plupart des algorithmes de reconnaissance de formes permettant de calculer la probabilité *a posteriori* grâce à la probabilité *a priori* :

$$P(w_t \mid u_t, a_t, s_t, n_t) = \frac{P(u_t \mid w_t, a_t, s_t, n_t) \cdot \overbrace{P(w_t \mid a_t, s_t, n_t)}^{\text{Modèle de Langage}}}{P(u_t \mid a_t, s_t, n_t)}$$

Notons que le terme $P(w_t \mid a_t, s_t, n_t)$ représente le modèle de langage (c'est à dire la probabilité d'occurrence d'un mot dans le langage étudié). On peut aussi simplifier ce terme puisqu'il est indépendant du bruit. Par contre, il reste conditionné par a_t et s_t car c'est une caractéristique importante des systèmes de dialogues que de pouvoir adapter le modèle de langage (et donc les séquences de mots autorisées) en fonction du contexte. En effet, afin d'améliorer les performances de reconnaissance, un système de dialogue peut décider de n'accepter que les phrases répondant à une question posée comme entrées valides et donc limiter le nombre de possibilités.

Génération de Parole

Enfin, le terme de génération de parole $P(t_t, i_t \mid a_t)$ peut lui aussi être décomposé de la manière suivante :

$$P(t_t, i_t \mid a_t, s_t) = P(i_t \mid t_t, a_t, s_t) \cdot P(t_t \mid a_t, s_t)$$
$$= \overbrace{P(i_t \mid t_t)}^{\text{TTS}} \cdot \overbrace{P(t_t \mid a_t, s_t)}^{\text{NLG}}$$

Le processus de synthèse de parole n'étant dépendant que du texte à synthétiser, le terme associé a été simplifié. Dans la plupart des systèmes de dialogues, il n'y a pas réellement de génération de langage naturel, c'est à dire de synthèse de texte à partir de concepts [6]. Les textes sont le plus généralement écrits par les concepteurs du système pour être ensuite synthétisés (parfois même, la synthèse vocale est absente et le système utilise des fichiers audio préenregistrés). Néanmoins, le développement de modules NLG dans le cadre de systèmes de dialogues prend de l'ampleur actuellement [8] et le résultat n'est donc pas toujours déterministe, ce qui justifie la description probabiliste qui est aussi réalisée pour ce procédé. On peut dire que le processus de génération de texte est dépendant de l'état dans lequel se trouve le système car, suivant l'historique, on pourra par exemple pronominaliser certains sujets ou compléments s'ils ont déjà été mentionnés plus tôt dans le dialogue ou réaliser des anaphores.

UTILISATION DU CADRE PROBABILISTE

Avec cette description probabiliste de la communication parlée entre homme et machine, il est possible de réaliser plusieurs applications utiles lors de la conception de systèmes de dialogues. Dans tous les cas, il sera nécessaire d'estimer en tout ou en partie les probabilités décrites dans les paragraphes précédents.

Estimation des Probabilités

L'avantage de la description probabiliste qui a été faite dans cet article est qu'elle permet de désolidariser les modules les uns des autres mais aussi de les découpler de la tâche pour certains d'entre eux. Ceci permet non seulement d'estimer les probabilités relatives à chacun des modules de manière individuelle mais aussi d'estimer ces probabilités sur des corpora de données qui ne sont pas forcément relatives à la tâche étudiée. Comme il est très souvent impossible de disposer d'un corpus de dialogue complètement relatif à la tâche, surtout en phase de développement, ceci est très pratique. De plus, les modèles estimés peuvent aussi parfois être réutilisés pour le développement d'autres systèmes de dialogues.

Simulation et Validation

Lors de la conception d'un système de dialogues il est souvent nécessaire de valider le comportement d'un ou de plusieurs modules alors que les autres ne sont pas disponibles. De plus, même si tous les éléments du systèmes sont disponibles, il est parfois difficile de

réaliser la validation d'un module particulier dans les conditions réelles d'utilisation parce que cela prend trop de temps et que cela ne peut pas se faire de manière automatique et systématique (par exemple, réaliser un très grand nombre de fois une reconnaissance de parole pour évaluer le système de compréhension de langage naturel prend beaucoup de temps). Dans ce cadre, remplacer un module par son modèle probabiliste peut s'avérer très utile. Dans [4], par exemple, les auteurs présentent un modèle d'utilisateur relativement simple dans le but d'évaluer le module de gestion du dialogue. Dans le même esprit, il est proposé dans [5] de modéliser un système de reconnaissance de parole.

Optimisation

Comme déjà mentionné auparavant, la conception du système de gestion de dialogue est un des points les plus critiques du développement d'une interface vocale puisque c'est lui qui est responsable de la coordination de l'interaction et qui influe donc terriblement sur l'ergonomie de l'interface. La possibilité de simuler automatiquement le comportement des différents modules entourant le gestionnaire de dialogue permet non seulement d'en évaluer les performances mais laisse aussi la porte ouverte à la possibilité d'apprentissage automatique de stratégies optimales. En effet, la description d'un dialogue comme un processus séquentiel stochastique s'inscrit parfaitement dans le cadre des processus de décision de Markov et de l'apprentissage non-supervisé par renforcement [7]. Ces techniques nécessitant un nombre important d'interactions pour converger, la possibilité de simuler de manière réaliste ces interactions, en tout ou en partie, constitue une perspective intéressante.

CONCLUSIONS ET PERSPECTIVES

Dans cet article, une description probabiliste de la communication parlée entre homme et machine a été exposée. Cette description est basée sur le parcours de l'information à travers les différents modules qui composent le système de dialogue. Grâce à cette description, une modélisation comportementale des différents modules ainsi que du système complet peut être envisagée par le biais de l'estimation de certains paramètres. Cette modélisation peut être utile dans le cadre de la validation, de l'évaluation et de l'optimisation automatique de la gestion de dialogue.

Un aspect du traitement de l'information par les différents modules d'un système de dialogues vocaux a été volontairement omis dans cet article, celui de la production de métriques par les différents modules apportant de l'information sur la qualité du traitement. En effet, la plupart des systèmes actuels de reconnaissance vocale, par exemple, sont capables de fournir un niveau de confiance accompagnant le résultat. Ces métriques sont aussi souvent utilisées par les systèmes de gestion de dialogue pour mettre à jour leur état interne. Elles peuvent aussi être utiles pour construire un critère d'évaluation de l'interaction qui peut ensuite être utilisé dans le cadre d'une optimisation automatique. Ces aspects doivent être intégrés dans la description probabiliste exposée dans cet article.

BIBLIOGRAPHIE

1. Allen, J. *Natural Language Understanding*. Benjamin Cummings, 1987, Second Edition, 1994.
2. Boite, R., Bourlard, H., Dutoit, T., Hancq, J., Leich H. *Traitement de la Parole*, 2nd Edition. Presses Polytechniques Universitaires Romandes, Lausanne, ISBN 2-88074-388-5, 2000.
3. Dutoit, T. *An Introduction to Text-To-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht, 320 pp., ISBN 0-7923-4498-7, 1997.
4. Eckert, W., Levin, E., Pieraccini, R. User Modeling for Spoken Dialogue System Evaluation. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'97)*, 1997, pp. 80-87.
5. Pietquin, O. and Dutoit, T. Modélisation d'un Système de Reconnaissance dans le Cadre de l'Evaluation et l'Optimisation Automatique des Systèmes de Dialogue. *Actes des 'Journées d'Etude de la Parole' (JEP'02)*, Nancy, France, Juin 2002.
6. Reiter, E. and Dale, R. *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, 2000.
7. Sutton, R. and Barto, A. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.
8. Walker, M., Rambow, O., Rogati, M. Training a Sentence Planner for Spoken Dialogue Using Boosting. In *Computer Speech and Language Special Issue on Spoken Language Generation*, July 2002.