A Probabilistic Description of Man-Machine Spoken Communication

Olivier Pietquin[†]

École Supérieure d'Électricité (Supélec)
Signal Processing Systems group
2 rue Edouard Belin
F-57070 Metz – France
olivier.pietquin@supelec.fr

Faculty of Engineering, Mons
TCTS Lab
1, av Copernic
B-7000 Mons – Belgium
pietquin@tcts.fpms.ac.be

Abstract

Speech enabled interfaces and spoken dialog systems are mostly based on statistical speech and language processing modules. Their behavior is therefore not deterministic and hardly predictable. This makes the simulation and the optimization of such systems performances difficult, as well as the reuse of previous work to build new systems. In the aim of a partially automated optimization of such systems, this paper presents a formalism attempt for the description of man-machine spoken communication in the framework of spoken dialog systems. This formalization is partly based on a probabilistic description of the information processing occurring in each module composing a spoken dialog system but also on a stochastic user modeling. Eventually, some possible applications of this theoretic framework are proposed.

1. 1. Introduction

Speech enabled interfaces are becoming more and more present in our day-to-day life. Automatic phone call routing or voice dialing for example are becoming ubiquitous. Yet, speech and language processing are techniques based on statistical methods, errors are possible and the behavior of voice-based interfaces is therefore hardly predictable. In this paper, we propose a probabilistic framework for the description of the man-machine spoken communication.

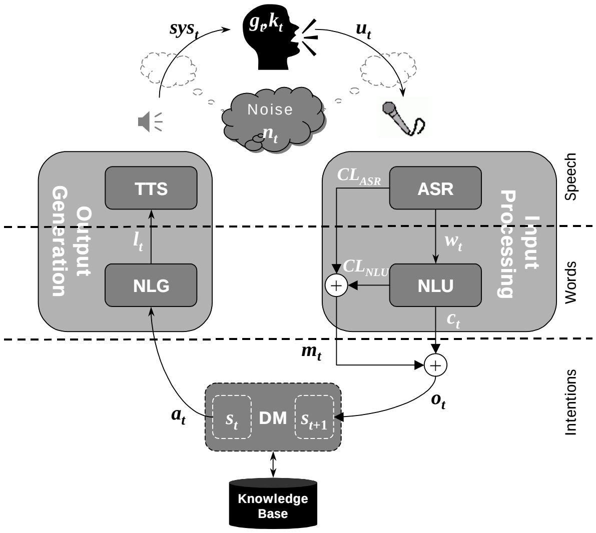

Figure 1: Man-Machine Spoken Dialog

To do so, a man-machine spoken dialog will be considered as a sequential process in which a human user and a Dialog Manager (DM) are communicating using speech through speech and language processing modules (Figure 1). The role of the DM is to define the sequencing of spoken interactions and therefore to take decisions about what to do at a given time (providing information, asking for information, closing the dialog, etc.). A Spoken Dialog System (SDS) is often meant to provide information to a user; this is why it is generally connected to a knowledge base through its DM. The dialog is therefore regarded as a turn-taking process in which pieces of information are processed sequentially by a set of modules and perform a cycle going from the DM to the user and back. At each turn $\boldsymbol{t}$ the DM generates a communicative act set $\boldsymbol{a_t}$ according to its internal state $\boldsymbol{s_t}$ and corresponding to its decision about what to do in that state. This act set is then transformed into a linguistic representation $\boldsymbol{l_t}$ (generally a text) by a Natural Language Processing (NLP) module. The textual representation l_t serves as an input to a

[†] This work was realized when the author was with the Faculty of Engineering, Mons (FPMs, Belgium) and was sponsored by the 'Direction Générale des Technologies, de la Recherche et de l'Énergie' (DGTRE) of the Walloon Region (Belgium, First Europe Convention n° 991/4351).

Text-to-Speech synthesizer (TTS) to produce a system spoken output ***sys**$_t$*. The TTS and the NLG modules are therefore spoken Output Generation modules. To this spoken solicitation, the user answers by a new spoken utterance $\boldsymbol{u_t}$ according to what he could understand from sys_t, to his/her knowledge $\boldsymbol{k_t}$ (about the task, the interaction history, the world in general) and to the goal $\boldsymbol{g_t}$ s/he is trying to achieve by interacting with the system. Both spoken utterances sys_t and u_t can be mixed with some noise $\boldsymbol{n_t}$. The noisy user utterance is in turn processed by an Automatic Speech Recognition system (ASR), which produces a written word sequence $\boldsymbol{w_t}$ as a result and a confidence measure $\boldsymbol{CL_{ASR}}$ about this result. A Natural Language Understanding module (NLU) subsequently tries to extract communicative acts (or concepts) $\boldsymbol{c_t}$ from w_t (possibly helped by CL_{ASR}). The NLU module also provides some confidence measure $\boldsymbol{CL_{NLU}}$ about this processing. The NLU and ASR sub-systems are Speech Input processing modules. The set $\{c_t, CL_{ASR}, CL_{NLU}\}$ composes an observation $\boldsymbol{o_t}$ which can be considered as the result of the processing of the DM communicative acts a_t by its *environment*. Therefore, the DM computes a new internal state $\boldsymbol{s_{t+1}}$ according to this observation.

The following paragraphs will use this description of a man-machine dialog as a base to build a probabilistic model.

2. Probabilistic Description

From the point of view of the DM, the interaction can probabilistically be described by the joint probability of the signals a_t, o_t and s_{t+1} given the history of the interaction:

$$\begin{aligned}
&P(s_{t+1},o_t,a_t \mid s_t,n_t,a_{t-1},s_{t-1},n_{t-1},\ldots,a_0,s_0,n_0) = \\
&\quad \underbrace{P(s_{t+1} \mid o_t,a_t,s_t,n_t,a_{t-1},s_{t-1},n_{t-1},\ldots,a_0,s_0,n_0)}_{\text{Task Model}} \cdot \\
&\underbrace{P(o_t \mid a_t,s_t,n_t,a_{t-1},s_{t-1},n_{t-1},\ldots,a_0,s_0,n_0)}_{\text{Environment}} \cdot \qquad (1) \\
&\underbrace{P(a_t \mid s_t,n_t,a_{t-1},s_{t-1},n_{t-1},\ldots,a_0,s_0,n_0)}_{\text{DM}}
\end{aligned}$$

In (1), the first term stands for the *task model* that helps building a new DM internal state thanks to the received observation, the second term stands for the response of the *environment* to the DM stimulation, and the third stands for the DM decision process.

2.1. Markov Property and Random Noise

In the case of a SDS, the Markov Property is met if the DM choice about the action a_t to perform at time t and its state s_{t+1} at time t+1 are only relying on the state at time t and not of previous states and actions. From now on, the Markov Property will be assumed. It can anyway be met by a judicious choice of the DM state representation, which should embed the history of the interaction into the current state. Such a state representation is said *informational*. The same assumption is made about the environment. Moreover, the noise will also be considered as being random so as to have independence between n_t and n_{t-1}. Eq. (1) then becomes:

$$\begin{aligned}
P(s_{t+1},o_t,a_t \mid s_t,n_t) &= \underbrace{P(s_{t+1} \mid o_t,a_t,s_t,n_t)}_{\text{Task Model}} \cdot \\
&\quad \underbrace{P(o_t \mid a_t,s_t,n_t)}_{\text{Environment.}} \cdot \underbrace{P(a_t \mid s_t,n_t)}_{\text{DM}} \qquad (2)
\end{aligned}$$

2.2. Environment

The term associated with the environment in (2) can also be written as:

$$\begin{aligned}
P(o \mid a,s,n) &= \sum_{sys,k,g,u} P(o,sys,k,g,u \mid a,s,n) \\
&= \sum_{sys,k,g,u} \underbrace{P(sys \mid a,s,n)}_{\text{Output Generation}} \cdot \underbrace{P(o \mid u,g,sys,a,s,n)}_{\text{Input Processing}} \cdot \\
&\qquad \underbrace{P(k \mid sys,a,s,n)}_{\text{Knowledge Update}} \cdot \underbrace{P(g \mid k,sys,a,s,n)}_{\text{Goal Modification}} \cdot \qquad (3) \\
&\qquad \underbrace{P(u \mid g,k,sys,a,s,n)}_{\text{User Utterance}}
\end{aligned}$$

where the t indices were omitted for the sake of clarity. This last equation links all the signals playing a role in the DM communicative acts processing by the environment. There are terms associated with the output generation and input processing modules but also other terms related to the user. Before going further, some more assumptions can be made in order to simplify (3):

- Although one could think about sound level adaptation, the actual spoken utterance *sys* is not dependent of noise. The perception of *sys* by the user can be degraded because of noise, but the generation process is not tuned according to it.
- The input processing is independent from the user's goal g of course, but also from *sys*. Indeed, if the DM decision can be responsible for some tuning (i.e. loading ASR grammars etc.), the spoken realization *sys* of the communicative acts is not responsible for any tuning of the input processing modules.
- The u signal does not rely on the DM communicative acts a and state s. Indeed, the communicative acts are transmitted to the user through *sys*.
- A goal modification can only occur because of a user's knowledge update that happens thanks to the last system utterance *sys*.
- The DM acts a don't influence the knowledge update. Indeed, from the user's point of view, they are

only conveyed by the last system utterance.

From this, (3) becomes:

$$\begin{aligned}&\sum_{sys,k,g,u} P(o,sys,k,g,u \mid a,s,n) = \\ &\sum_{sys,k,g,u} \underbrace{P(sys \mid a,s)}_{\text{Output Generation}} \cdot \underbrace{P(o \mid u,a,s,n)}_{\text{Input Processing}} \cdot \underbrace{P(k \mid sys,s,n)}_{\text{Knowledge Update}} \cdot \\ &\quad \underbrace{P(g \mid k)}_{\text{Goal Modification}} \cdot \underbrace{P(u \mid g,k,sys,n)}_{\text{User Utterance}}\end{aligned} \quad (4)$$

2.3. User

In (4), the terms associated to the user are:

$$\underbrace{P(k \mid sys,s,n)}_{\text{Knowledge Update}} \cdot \underbrace{P(g \mid k)}_{\text{Goal Modification}} \cdot \underbrace{P(u \mid g,k,sys,n)}_{\text{User Utterance}}. \quad (5)$$

They express the existing relation between the user's speech production process and his/her goal and knowledge, themselves related to each other. Besides, the knowledge can be modified at any time through a system utterance. Yet, this modification of the knowledge is incremental and depends on the last system utterance (which might be misunderstood, and especially in presence of noise) and the previous user's knowledge state. Thus, one can also write:

$$\begin{aligned}P(k \mid sys,s,n) &= \sum_{k^-} P(k \mid k^-,sys,s,n) \cdot P(k^- \mid sys,s,n) \\ &= \sum_{k^-} P(k \mid k^-,sys,n) \cdot P(k^- \mid s),\end{aligned} \quad (6)$$

where ***k***$^-$ stands for k_{t-1}. The simplifications in (6) come from the fact that the current noise (since it is supposed to be random) cannot influence the correspondence between the previous knowledge and system state. In the same way, the current system utterance doesn't affect the previous knowledge state. This relation between the update of the user's knowledge and the system state can be easily related to the grounding process occurring during human-human dialogs [1]. From (6) one can notice that, similarly to the SDS, which relies on a state update, the user's behavior relies on his/her knowledge update. Yet the system and the user do not rely on the same state space and this is the source of possible misunderstandings between the user and the system and *vice-versa*.

2.4. Input Processing

The term related to the input processing in (4) can also be factored as:

$$\begin{aligned}P(o \mid u,a,s,n) &= P(c,CL_{ASR},CL_{NLU} \mid u,a,s,n) \\ &= \sum_w P(c,CL_{ASR},CL_{NLU} \mid w,u,a,s,n) \cdot P(w \mid u,a,s,n) \\ &= \sum_w P(c,CL_{NLU} \mid w,CL_{ASR},u,a,s,n) \cdot P(w,CL_{ASR} \mid u,a,s,n) \\ &= \sum_w P(c,CL_{NLU} \mid w,CL_{ASR},a,s) \cdot P(w,CL_{ASR} \mid u,a,s,n) \\ &\approx \max_w \underbrace{P(c,CL_{NLU} \mid w,CL_{ASR},a,s)}_{\text{NLU}} \cdot \underbrace{P(w,CL_{ASR} \mid u,a,s,n)}_{\text{ASR}}.\end{aligned} \quad (7)$$

In (7), the last equality is obtained by assuming that the NLU process doesn't rely on the acoustics but only on the ASR results, which result themselves from a maximization process.

2.5. Output Generation

The output generation subsystems include the NLG and the TTS modules. The output generation term of (4) can therefore be factored as:

$$\begin{aligned}\underbrace{P(sys \mid a,s)}_{\text{Output Gen.}} &= \sum_l P(sys \mid l,a,s) \cdot P(l \mid a,s) \\ &= \sum_l \underbrace{P(sys \mid l)}_{\text{TTS}} \cdot \underbrace{P(l \mid a,s)}_{\text{NLG}} \quad .\end{aligned} \quad (8)$$

Here, it is assumed that the DM act and state don't influence the TTS process but only the text *l* generated by the NLG module. Generally, one text *l* corresponds to only one synthesized utterances *sys* so the summation over all possible texts in (8) is not required.

3. Possible Uses of the Framework

This probabilistic framework helps in identifying the different stochastic variables influencing the functioning of SDS modules. It can be used in several applications related to SDS design such as strategy evaluation and validation or optimal strategy learning. Indeed, a model showing the same probabilistic properties could replace each SDS module for simulation purpose. To do so, a way to estimate the probabilities occurring in the different abovementioned equations is required.

3.1. Probability Estimates

The development exposed in section 2 allows disuniting the probabilities related to the different modules composing a SDS. In a certain extent, it also allows building task-independent models for some particular modules like the ASR system for example, as proposed in [2]. This method could be used to estimate the ASR related term of (7):

$$\begin{aligned}P(w,CL_{ASR} \mid u,a,s,n) = P(w \mid u,a,s,n) \cdot \\ P(CL_{ASR} \mid w,u,a,s,n)\end{aligned} \quad (9)$$

Indeed, the couple (a,s) often determines the contextual Language Model (LM) used for speech recognition and therefore, a method for estimating ASR performances according to a given LM is suitable.

On another hand, the user behavior described by (5) and (6) implies different probabilities that are quite difficult to estimate from experience or corpora. Yet it could be modeled thanks to simpler probabilities like proposed in [3] where a Bayesian-Network-based (BN-based) approach to user modeling is proposed. Therefore, other simpler probabilities could be estimated on corpora or assessed by experts. This BN-based user model is also used as a classification tool to simulate the NLU part of (7):

$$P(c, CL_{NLU} \mid w, CL_{ASR}, a, s) = P(CL_{NLU} \mid c, w, CL_{ASR}, a, s) \cdot P(c \mid w, CL_{ASR}, a, s) \quad (10)$$

Finally, the NLG term of (8) should be used to model errors in the transmission of the concepts embedded in *a*. This generally occurs because of bad references in the generation of pronouns or anaphora. Therefore NLG errors only arise when referring to already referred subjects. It is proposed to add an ambiguity parameter $\xi \in [0,1]$ modifying the meaning of a *sys* when *a* refers to a subject already mentioned in the conversation. If ξ is non zero, *a* might be mismatched with another *a´*:

$$P(l = f(a) \mid a, s) = 1 - \xi \text{ and } P(l = f(a') \mid a, s) = \xi \quad (11)$$

This mismatch results generally in a user misunderstanding but it is due to an ambiguity in the generated sentence.

3.2. Simulation, Validation and Optimization

During the design process of a SDS, it can happen that some modules have to be tested and validated while the others are simply not available. For example the dialog manager can be evaluated without a human user or all the other artificial modules. Moreover, the validation of a SDS in real conditions is time-consuming. Thus, replacing some modules by their probabilistic model is desirable. There exist previous examples of dialog simulation for evaluation purposes like in [4] where it is proposed to model the user's behavior to rapidly evaluate the quality of a dialog manager strategy.

The substitution of one or several modules by their probabilistic model can also be interesting for automatic learning of optimal dialog strategies by means of unsupervised learning techniques like initially proposed in [5]. Results of the application of the exposed probabilistic framework to the problem of dialog management optimization can be found in [3].

4. Conclusion and Perspectives

In this paper, a probabilistic framework for the description of man-machine spoken communication is described. It is based on the processing cycles of the information by the different modules composing a Spoken Dialog System and by the human user. From this framework, a behavioral modeling of each submodule is possible as well as at a global level thanks to several parameter estimations. Some ways to estimate those parameters have been given and some possible uses of this framework, such as evaluation and automatic optimization of SDS subsystems were also proposed.

In the future, man-machine interfaces will probably rely on multimodal communication and not only on spoken communication anymore. Stochastic modeling of such interfaces could help in designing more user-friendly interfaces according to some objective and subjective criterion by enabling prior testing and evaluation. The description exposed in this paper could be a starting point for a more general man-machine communication formal description.

5. References

[1] H. Clark, E. Schaefer, '*Contributing to Discourse*.' Cognitive Science, 13, 1989, pp.259-294.

[2] O. Pietquin, *A Framework for Unsupervised Learning of Dialogue Strategies*, Presses Universitaires de Louvain, SIMILAR Collection, ISBN 2-930344-63-6, 2004, pp. 158-167.

[3] O. Pietquin, T. Dutoit, '*A Probabilistic Framework for Dialog Simulation and Optimal Strategy Learning*' IEEE Transaction on Speech and Audio Processing, Spec. Issue on Data Mining of Speech, Audio and Dialog. *To appear*.

[4] Eckert, W., Levin, E., Pieraccini, R. '*User Modeling for Spoken Dialogue System Evaluation.*' Proc. ASRU'97, 1997, pp. 80-87.

[5] E. Levin, R. Pieraccini, '*A Stochastic Model of Computer-Human Interaction for Learning Dialogue Strategies,'* Proc. Eurospeech'97, Rhodes, Greece, 1997, pp. 1883-1886.