Three Steps
Towards Real Artificial Speech Communication

Mark Liberman

University of Pennsylvania
http://ling.upenn.edu/~myl

ABSTRACT: Three Steps Towards Real Artificial Speech Communication

1. Robust diarization, i.e. "who spoke when". Despite decades of effort, the best diarization systems have trouble with overlapping speech and noisy backgrounds – i.e. with real life.

2. Dialogue systems with human-like turn-taking. People break into others' turns for many reasons, cooperative as well as competitive. Dialogue systems typically interpret interruptions as reset signals, which they usually aren't. And systems generally don't speak during an interlocutor's turn, not even for backchannels, much less to correct misunderstandings, introduce relevant information, or cut to a proposed solution.

3. Conversational systems that can participate effectively in a meeting, manage a classroom, or chat over a game of poker. Such systems need to keep track of what is happening in the physical and social context, as well as who said what when. They need to learn how and when to contribute, and how to modulate their contributions dynamically as a function of others' uptake (or lack thereof). And they need to be able to adapt to different conversational cultures.

Sally Jacoby & Elinor Ochs, “Co-Construction: An Introduction”, 1995:

“The primary contribution of ethnomethodology and conversation analysis has been to demonstrate that: social interaction is itself an exquisite accomplishment.

Almost three decades ago, Garfinkel (1967) introduced the idea that familiar and unproblematic as they may appear, mundane social encounters rely on detailed indexical understandings of what might be happening right now, what just happened, and what will likely happen next in some particular, located routine activity.

An important idea in this research is that actions are accomplished and utterances understood crucially because others are filling in common-sense understandings entailed in the situation at hand.

That is, sensemaking is an interactional affair.”

Examples are everywhere.

Here’s a dinner party recording
from the *Santa Barbara Corpus of Spoken English*, *Part 1*:

In everyday life,
we have no trouble sorting out who said what when,
despite overlapping speech, background noise, changes in voice quality,…

In Speech Technology World,
figuring out ***who spoke when*** is called *"diarization"* –
and the best of today's systems can't deal with a typical dinner party –
or even a political debate (which should be easier)…

Robust diarization –
who spoke when, understood in real time as the interaction unfolds –
is the first thing a system needs
in order to think about becoming a conversational participant.

And robust speech recognition in such cases –
who spoke **what** when –
is harder, and further from today's state of the art.

But even just for monitoring human interaction,
we need to start with simple diarization.

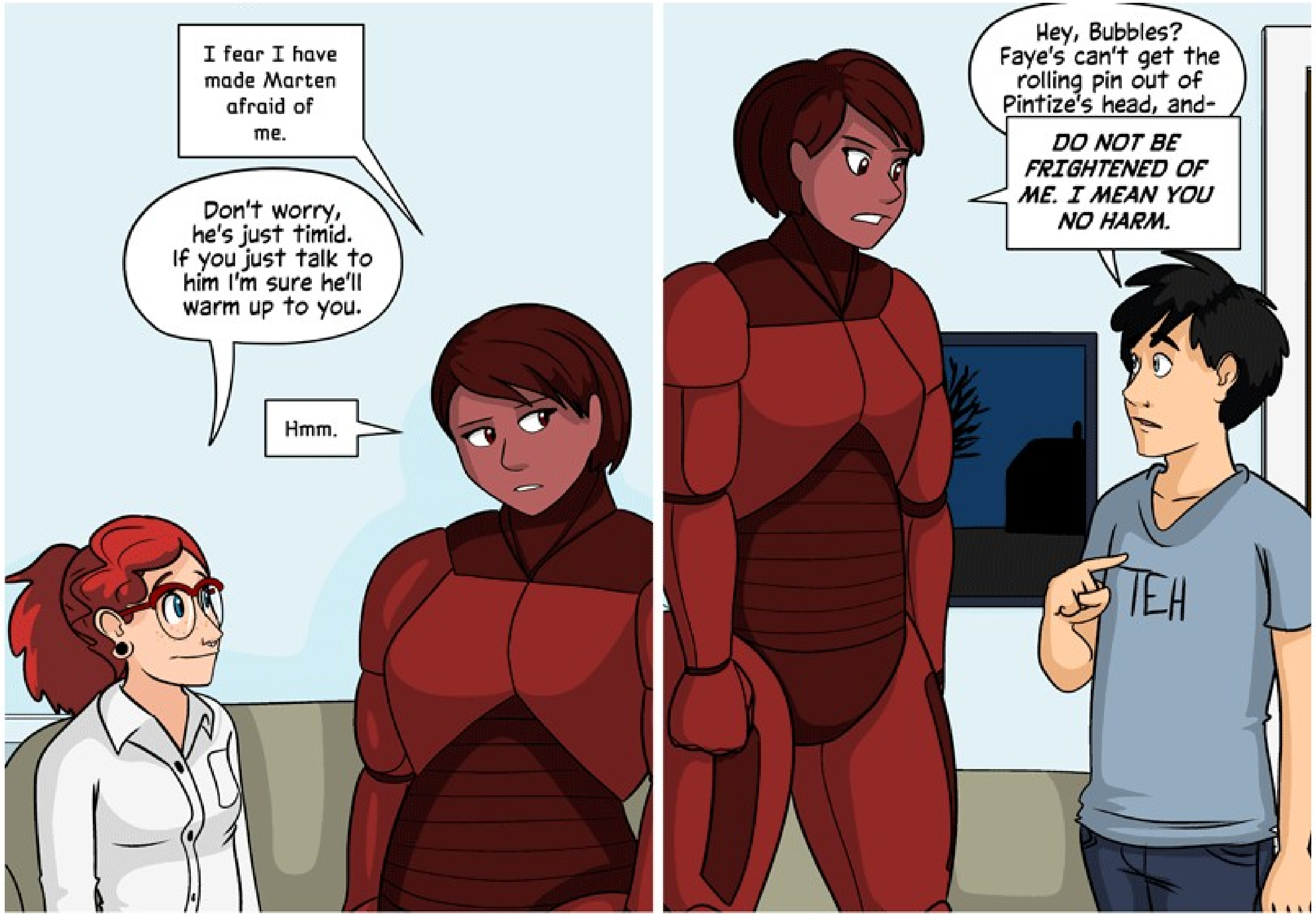
Robust diarization has intrinsic value
as an indicator of personality, mood, engagement, relationships
in clinical and educational settings –
and partial solutions are within reach.

Further progress will require some mixture
of smarter signal processing,
better recognition of the content of speech,
and deeper modeling of patterns of interaction.

It isn’t clear how much of which we need --

but it **IS** clear that diarization is not yet a solved problem.

For more than half a century,
we’ve imagined systems that don’t just monitor human interactions,
but actually take part in them…


I fear I have made Marten afraid of me.
Don't worry, he's just timid. If you just talk to him I'm sure he'll warm up to you.
Hmm.
Hey, Bubbles? Faye's can't get the rolling pin out of Pintize's head, and-
DO NOT BE FRIGHTENED OF ME. I MEAN YOU NO HARM.
TEH

Engineers usually frame the goal as overtly useful communication,
like intelligent tutoring, interactive travel advice, or

But Siri’s initial popularity was partly driven by impractical ”sassy” repartee.

A system that could add entertaining and appropriate comments
to conversations at the dinner table or the entertainment console
might be similarly be a hit –

And it certainly would let us explore models of interactional sense-making.

The need to participate from a distance in today's workshop
suggested something completely different
that better models of communicative interaction
might be good for…

I'm sorry to miss this evening's informal post-workshop get-together –

among the many things that telepresence is really bad at,
events like that are near the top of the list.

We can imagine an app that facilitates
virtual distributed dinner parties,
whereby people in restaurants and dining rooms all around the world
could enjoy one another’s company
as effectively as if they were all sitting at the same table.

But that app would need to have a good model
of the dynamics of conversational interaction,
and would need to be able to apply it in real time
so as to let interacting groups of different sizes
form and dissolve In a natural way.

And, of course, the artificially intelligent host would perform introductions,
fill awkward silences by introducing or reviving appropriate topics,
and gracefully rearrange the table from time to time
so as to engage people who might otherwise be left out …

Thanks for your attention!

?