# Captions

## (M1D)

## The Sign2 Project: Translation of American Sign Language to Audio and Text

**Chance Glenn**

ROUGH EDITED COPY

NTID/RIT

THE SIGN2 PROJECT: TRANSLATION OF AMERICAN SIGN LANGUAGE TO AUDIO AND TEXT

CHANCE GLENN

JUNE 27, 2005

1:00 P.M. CST

CAPTIONING PROVIDED BY: CAPTION FIRST, INC.

PO BOX 1924

LOMBARD, IL 60148

* * * * *

This is being provided in a rough-draft format. Communication Access Realtime Translation (CART) is provided in order to facilitate communication accessibility and may not be a totally verbatim record of the proceedings

* * * *

>> It's 1:03, and I think we're ready to start.

Good afternoon.

We have a wonderful presenter this afternoon.

He will introduce himself.

But I am here to ask to you fill out the evaluation form.

It's the yellow paper that I will hand out to you.

And then at the end of the presentation, if you would fill out that evaluation form and hand it back to me.

Enjoy the rest of the afternoon.

>> CHANCE GLENN: Hello, everybody.

How are you? I represent the laboratory for advanced communications technology here at RIT as well as the -- well, the laboratory is apart of the center for advancing the study of cyber infrastructure.

And that's a mouthful.

So we just call it CASCI.

I am part of the telecommunications engineering technology department here at RIT.

So in that, I've been here a couple of years.

I'm going in my third year here at RIT.

And we've been developing technology that we believe will be beneficial to bridging the communication gap exists.

I know that I can speak specifically about that gap because I don't sign.

And so I would not be able to directly communicate with someone who only signs.

And so that's an issue for me.

One of the things that occurred to me was being here at RIT which is one of the leading technology schools in the country, and also RIT having the national technical institute for the deaf here which is doing world-class and leading research in aid for technology for the deaf, as well as other areas of research.

It makes sense that we should also lead in technical development in the same way that some other places are doing.

And so I have some ideas, and I began to talk with some colleagues of mine, and this is what we've come about.

Some of the others that are mentioned here are students of mine that have been working on this project.

I just wanted to take a moment and acknowledge them.

Well, what is sign2?

I use that name means Sign Language to something else.

That's the goal.

And so specifically this is a focused research and development effort that has an end toward being a prototype and a technical device that could be used to translate Sign Language to either audio or text via computer or image processing.

The threefold goal is this: One is to further establish and enhance the body of knowledge in physical movement and its translation to language.

Number two is to conceptualize and engineer a prototype that closes the communication gap, or bridges the communication gap between the deaf and the hearing.

And thirdly, to establish and build a statistical database from the prototype that we engineer that has results that are useful to research and development worldwide so we want to take the things that we develop and disseminate them throughout the community for the benefit of all that would need it.

Well, here's the purpose.

I just use these images.

Some of these came right off of MicroSoft.

But anyway, the idea is to bridge this communication gap.

Now, say that this person is deaf and able to sign, and this person, like me, is unable to sign.

Or unable to read Sign Language.

So this is the direction that we want to deal with.

And it so happens that in this image this person is holding an electronic device.

It could be like a PDA or a smart phone or other technical devices that would actually image the Sign Language that is occurring here and translate it into text or audio for the person here.

That's the idea.

Now, there are other methods that are being discussed at this symposium as well as other technical forums that have dealt with the communication direction from this way (indicating) to this way.

And we're dealing with this direction (indicating).

So this is, as you might imagine, a very daunting undertaking.

The human brain is a very amazing and complex thing.

To try to mimic the processes that the brain undertakes is challenging, to say the least, probably if not impossible.

But yet to try to take a natural approach because what we do when we're signing, you are using your eyes to see what the signer is producing, and then translating it into your brain into language.

So a natural consequence of how we would want to approach it is to use the same method, to use an imaging system, or a visual system to do this.

There has been other technical approaches, for example, using data gloves.

Putting gloves that can pinpoint the positions of the hand and the arms and things of that nature, and then using that information to translate that into positions, and, therefore, translate that into language.

But it seems to me that that would be, number one, very cumbersome, and, number two, very intrusive particularly to the signer.

How would it be for me to say, well, I want to communicate with you.

Here, put these gloves on and then communicate.

That's not the way that we want to go about it.

So we want a passive system, a system that does not require that the signer do anything but sign.

And then the technology takes it from there.

So this is the reasoning behind this image processing approach as opposed to other types of things.

Another reason -- and these are the reasons.

As I said, it's more of a natural approach.

It's less intrusive to the signer.

Also, there are data reduction techniques that are already available in the form of image compression, feature extraction, those types of things can be done with images, less so with other types of technology that are less reliable.

Again, image processing techniques can be integrated with already-standing technology such as PDAs, as I mentioned, smart phones, video phones, kiosks, high-tech kiosks.

You could envision in a mall or at any kind of service center, you could come up and sign and it could be translated directly into text for someone to meet your needs.

So this is the idea.

And this is the reasoning.

We're doing this in phases.

Well, here is the concept, first of all.

We have this imaging device here, and this is the person that's signing, and there would be some storage in that.

And this is the block diagram that covers how the actual system would work.

First of all, you would have a capture device which would be a camera.

Then you are storing it.

And then there is processing for extraction.

And I will show you a little bit of that in a second.

And then the actual processing itself, and I will show you some of that.

And then there is a comparison to our statistical database.

I will talk about that in a little bit.

And due to this comparison to the database, you get an error, and then you do error correlation.

Depending on how the errors fall out, that's how you determine what letter or word has been formulated.

I will show you that in a second.

>> CHANCE GLENN: Okay.

As I mentioned, this is a tremendous undertaking, and we're not underestimating the complexity of this problem at all.

So what we want to do is we start with a simple type of undertaking and build outward from that.

So finger spelling would be where we start.

And to explain why, what would you envision is that you are taking video data off of a stream, and this would be video of someone fingerspelling a word, and you would extract the image at the proper frame position, and then this image is then processed, and I am going to demonstrate that right now for you.

I'm going to demonstrate how this actually works.

Hopefully I won't mess things up here.

So what I want to do is pull up one of the programs that we use.

Okay.

You what are seeing here is one of the extracted images from the video stream of someone signing the letter.

Obviously this is the letter "L." okay.

So it's interesting, you can just look at it and tell what it is.

But to make a computer say what that is is a little bit difficult.

Well, what you can do is you take this color image, and you can break it up into red, green, and blue components.

Now, if you notice because we used the blue background, see, blue drops out nearly completely from the red image, okay?

So this is the one that we use because we want to eliminate the background and only be left with this shape, okay?

Now, we discard these and keep this (indicating).

Once we have that, we can turn it more into a simple black and white by using a threshold, and basically say if you have a number above a threshold keep it, any numbers below the threshold throw it out.

So that's what you have right here (indicating).

Okay.

Now, the next thing I want to show you -- now, different people have different sizes of hands, different shapes.

All these different things as variations that you need to account for.

Also, your imaging device may be at different position at a different time.

So you have to accommodate for that as well.

And this is how we do it.

First of all, I want to show you something.

I'm going to change the threshold.

You see this white space here (indicating) that comes out?

If we adjust the threshold, we can eliminate that.

I am going to do that now, I hope.

>> CHANCE GLENN: So you see by changing the threshold a bit, I got rid of a lot of that, and it makes it more available for our next levels of processing.

So we eliminated the colors in the background as I showed you before.

Now you are left with this.

Now, in order to accommodate the different sizes and shapes of hands and things like that, what we need to do is get rid of the edges.

So this is where we eliminate the left edge, and basically just get rid of all of the white space here until we've reached this.

And then the next step is to crop the top edge, and then finally to crop the right edge.

So you are left with this shape right here (indicating).

So that's the processing procedure on one frame that is extracted from the video.

Now, if you want to see the actual process going through and actually see words coming out, we have a poster session at 3:00.

It will be set up, and we'll demonstrate for you a few of the words that we're able to recognize from this system.

We've had some pretty nice breakthroughs this year.

Okay, let me go back to where I was.

>> CHANCE GLENN: So once you have this processed image, then we apply a standard error calculation technique to calculate the mean square error from the subject or input image, and this is a statistical database of images that we store for the proper letter.

Now, in this database -- let me use some type of a pointer.

In this database, we have different subjects.

Here is subject 1, 2, 3, 4, to a large -- and we want to have a large number of subjects stored with processed images.

These images will be processed and stored on our computer.

That's how it's working now.

>From these subjects you have the different images that are associated with the proper letter.

Now, I know you what are saying.

There are a couple of letters that are movements, right?

But we use the initial position for, example, the letter "J" as it is stored in here.

And so we have that for each subject, and we have a number of subjects.

And also this database actually can be three-dimensional in that we have several instances of the same subject doing the same letter more than once.

That will be stored in our database.

Now, what we do is we take this input image from the video stream that we're processing, and we compare it to the images in the database.

This will give us an error according to this formula.

So this would be the original image.

This would be an image in the database (indicating).

We compare that, and we store the errors in this matrix, okay?

So then what we do is we sum the errors horizontally, and get an error sum according to this.

As you might imagine, the row that has the lowest error sum is the highest probability of that letter being formulated.

Okay?

So that's the process.

Pretty simple.

Nothing magic about it.

And here is a picture of the user interface that we have that actually we use to do this.

I'm going to explain what this is in a second.

But this is a video, this is running video of someone fingerspelling a particular word.

And here is some data that we need from that.

And here is the extracted images that come out at the right time (indicating).

That's what's used to determine the letter that's being formed at the particular time.

Now, let me talk about this for a second.

So how do you know when a letter is being formed?

Well, what we've been doing has been in a controlled situation.

We have a modest type of extraction system that we use in a room.

It's all nice and set up and the person is at a particular distance from the camera, and everything is controlled.

However, given that situation we have to start at some place.

So given that situation, we also have the person fingerspelling a little bit slower than would be natural.

But, again, if you like to see it, we'll have a demonstration down in the poster session.

So what we look at and what you see on this blue curve here is the errors between frames.

If we take a video of someone signing letters, we take every frame and we subtract it from the frame that just recently occurred.

This is the error that you get.

You can't see the numbers on the side here, but these numbers just range from 0-3.

So this means that there's not much difference between the frames and the video.

So what happens here?

What's going on here?

These are transitions between letters, okay?

So you see this jumping up here (indicating).

So that says that there is a transition.

So when you detect a transition, a good place to assume that a letter is in the middle of those transitions, okay?

And that's where we have extracted -- and this is where we choose to extract the frame say "Here's a letter."

So at this frame around 3 or so, there is a frame -- 340 or so, there is a frame that's a "Y." 210 there is "A."

And at 60-something there is a "L," okay?

So that's how we do that.

And this works under the conditions that we are at the present time.

So the process is the extraction of the frame itself and determining where that extraction should occur.

Then as I showed you before, the thresholding which allows you to create these black-and-white images.

And then the scaling, and then a comparison to the database which determines what letter has been formulated.

>> CHANCE GLENN: Okay.

So given what you've seen, and I keep advertising the poster presentation so I hope that you come down and see that.

I should have one of my students there demonstrating the work.

What do we need to do to really make this a reality, make it useful, things like that?

Well, first of all, we need to expand this database.

It is our contention that the more subjects we have in our database for comparison, the higher the probability of finding proper matches for a wide variation of people.

That's what we would like to do.

It's fine to say that it works for one person, but if it only works for one person that's just not going to get it done.

A lot like the technology that exists for speech-to-computer -- or voice recognition for computers.

You know, it's still not an exact science because different voices, different vocalizations, things like that.

That same type of thing occurs in signing as well.

We want to increase the reliability of letter extraction, and the determination process itself, meaning that we want to be able to sign -- have someone sign the same word over and over and over again and for it to be able to correctly determine the word that's been signed over a statistically large period of time.

That's the kind of work that we need to do and also for a large array of subjects.

We also need to improve the algorithm for more natural environments.

Right now it's not realistic to have someone, you know, in front of a blue screen.

We want to be able to have a system that could be fielded anywhere, and you can determine what someone is communicating.

Now, in the long term this is the real tough part here.

How do we extend this process for true signing?

With signing, it's more than just hand movements.

It's facial expressions, arm movements, all of these things are incorporated.

We believe that this process of extracting images is still prudent.

However, instead of analyzing one image for signing we would have to analyze a series of images and match them with a series of images that know formulated particular phrase in Sign Language.

So that's the goal.

That's how we're proceeding with this.

We also really want to develop a deeper working relationship with NTID, and particularly the deaf research community, in order to extend the connections to find out more about what needs to be done.

If we're going to develop a technology, how do you need to see it happen?

We don't want to come and say, "We have all of the answers this is how it's supposed to be done."

We want a back-and-forth communication on making this work the best.

Thirdly, we want to broaden the subject base for analysis, development for extending the current database that we have to broaden a database for full signing to formulate a proper format for the database for full signs.

Those are things that we need to do and we also want to develop a proper embodiment that will best aid in education, general communication.

What's the best way to put this out?

Mean, I have ideas.

People that work with me have ideas.

But we need to interact a little more to find out what's the best way to embody this technology if it's successful.

Okay.

Let me conclude by saying this: We developed this system to distinguish ASL fingerspelling using a purely image processing approach.

We've used this to resolve several words from various subjects so far with a high degree of reliability.

And we're developing this process for full-fledged ASL, and it's no reason why we couldn't extend this to other languages of signing as well.

That's certainly our goal.

We really want to provide a technology that's going to help reach the communication gap.

Of course, one more plug for that, and hopefully I can get all of that set up.

I haven't seen my student yet, and we're supposed to be set up at 3:00.

So hopefully all of that will be together.

I'd like to acknowledge a couple of people and organizations who have helped.

Of course, my department.

And I have some of our master students working on the project.

Also, undergraduate student who has done quite a bit of remarkable work on this.

I am proud of that.

Also for the center for supporting us, facilities, and some financial support.

Northstar Center has provided us student funding for some undergraduate researchers who have worked with us.

And also some early collaboration with NTID, and I want to, as I said, extend that.

And finally, here are references.

And I believe that these slides will be made available either online.

There is a paper that we put together that should be available after the conference as well.

You should have access to all of that.

You can easily find me under the department and e-mail me, write me or whatever you want to do to communicate, and I will be happy to hear from you.

I'm a leaving a little bit of time for any questions that you might have.

I am happy to entertain them now.

>> Audience member: What is object -- when you were saying about object, is it like different hands?

The object.

>> CHANCE GLENN: Let's see.

Can I go back?

Tell me where.

>> Audience member: He is asking about this.

What is this "S," you said it was object?

>> CHANCE GLENN: Subjects.

Different people.

So there would be different people with different hands who are signing the same thing.

>> Audience member: There is some research that's existing about using the idea for construction of the face by the photograph.

Do you think that this kind of idea could be applied to your topic?

>> CHANCE GLENN: Absolutely, yes.

There is one researcher, for example, his name is LijanÝYen, who is down in Binghamton University.

He has done quite a bit of work with feature extraction.

I know you can't see that probably, but facial animation using -- so we're working together, and we're looking at actually doing that.

>> Audience member: What do you think about the speed of computer?

How computers should increase if you can use this technology during like presentation or something like interpreters are doing, and the same speaker.

What do you think?

Your observation?

>> CHANCE GLENN: Well, first of all, if we use some of the extraction techniques that's mentioned here, as well as some other things, you can reduce the data that's

being processed.

So then you can process it faster.

Even now we are working in pseudo or almost realtime.

Not right yet.

But we're almost realtime.

So the computing power continues to go up and that's going to be to our advantage.

So we do think that it's realistic.

>> Audience member: I would like to communicate with you about publications, and bibliographies in terms of this topic.

>> CHANCE GLENN: I have a card.

>> Audience member: Thank you.

>> Audience member: I enjoyed your presentation.

One of my artificial intelligence classes did a paper along these lines.

One thing that kept coming up was that even no matter what you used, the gloves, that type of recognition, or the camera model, that even once you recognize the signs, unless it's in signed English, you still have the issue of then translating it into ASL.

I was wondering if you had any -- I know that's long term, but ideas of what you are going to do in terms of that second component?

>> CHANCE GLENN: Well, what you are talking about really is a mapping from ASL to English.

There is a lot of work that's been done on doing that.

So if we can take -- if we map our -- say for instance with Sign Language we would use series of images, okay?

So we take snapshots of how she is signing.

We take about four or however many.

We compare that to a database, there will be a ASL phrase that goes with it.

It may not be spoken English.

So then there would be a mapping from that ASL phrasing to English, and that's how we would try to do it.

Audience member: Thank you.

>> CHANCE GLENN: That's the goal anyway.

>> Audience member: I was wondering what the size of your vocabulary was, and what your accuracy rates are.

>> CHANCE GLENN: There are two answers to that.

Number one, we're using right now just fingerspelling.

So then we're using a comparison of all of the letters.

That's what's in the database.

I guess what you mean by the vocabulary is how many words we're able to distinguish at this point?

>> Audience member: It makes sense if it's 26.

>> CHANCE GLENN: That's the matching.

But the second part of your question almost helps with the first one, because I would think of the vocabulary being how many words can we successfully distinguish?

So far we have in the 10 to 20 range of different words.

And with the accuracy -- and we're still pulling numbers.

's not comfortable giving statistical numbers to the accuracy yet because we haven't done hundreds of tests yet, which is what I think that we should do.

But so far we've had a high degree in the above 80%.

So we want to push that higher, but also make that statistically meaningful when we say it.

>> Audience member: Before you mentioned in that grid the errors.

Can you identify what an error is?

>> CHANCE GLENN: Oh, yeah.

Let me just go back to it.

Think of this database being filled with images.

These images are the ones that represent that letter for this particular subject.

This is the one that we just extracted from our video that we wanted to distinguish.

So we take this and compare it to that (indicating), and actually a pixel-by-pixel comparison.

I just did it a second ago.

And so we make it the same size.

That's what the cropping and all of that that I showed before, it makes it the same size.

A point-by-point comparison sums up the differences and spits out a number.

The one -- if you obviously compared this image with itself (indicating), it will be zero.

So that means that's the one, right?

But if you have one in here that's very close to that it's going to be very close to zero.

And so what we're assuming is that different subjects even though they're different, the letters are going to be -- the images are going to be similar for the same letter.

So then you would get a bunch of low numbers if the letter was "B", okay?

>> Audience member: You made me think of a second question.

The letter "L" has a lot of form to, it but the letter "S" in your silhouette would just become a fist.

What do you do with that?

>> CHANCE GLENN: Well, there are not many like it, so we really -- if there are no other silhouettes -- so that difference, that simple difference between that and that (indicating), is enough to get the lower error sums.

So we pick the lower error sum.

That's why the statistical thing of getting more subjects will help it, okay?

So that's the idea.

Yes?

>> Audience member: In Hong Kong Sign Language, the "Z" is movement.

How do you recognize that?

>> CHANCE GLENN: To be accurate with that, we would have to use the approach that we're planning for full signing, and that's actually compare a series of images as opposed to just one.

Right now we're only taking the first formulation.

So in American Sign Language that's "J," and "Z," so we would take either the first position or the last position, or maybe somewhere in between, and just compare that.

That's not going to give us a good match.

So if you think about it, say that three images was enough to determine that your position, position, position, right?

All of these instead of just being one image it could be three.

Now, these would all be the same because there is no movement with these letters, but the "J" and the "Z" would have movement, and in those comparisons you would get the match.

So to answer your question, the movement would require more image comparison.

>> Audience member: What about the problem of left-hand side, right-hand side?

>> CHANCE GLENN: We haven't addressed that.

That's a good question.

These can be easily flipped, and so if you need to distinguish left hand or right hand you could have a second database that's everything flipped around and do that comparison, too, and determine through some process whether you are left or right-handed.

We just have not addressed that.

That's an idea just off the top of my head really.

I appreciate the questions.

These are good questions.

>> Thank you!

Remember to fill out the evaluation forms and give them to me.

Thank you.

* * * * *

This is being provided in a rough-draft format. Communication Access Realtime Translation (CART) is provided in order to facilitate communication accessibility and may not be a totally verbatim record of the proceedings

* * * *

Close